# A NOVEL BAP28 GENE AND PROTEIN

## Related Applications

The present application claims priority toU.S. Provisional Patent Application Serial No. 60/141,323, filed June 25, 1999 and U.S. Provisional Patent Application Serial No. 60/176,880, filed January 18, 2000, the disclosures of which are incorporated herein by reference in their entireties.

## FIELD OF THE INVENTION

The present invention is directed to polynucleotides encoding a human BAP28 polypeptide as well as a regulatory regions located at the 5'- and 3'-ends of said coding region. The invention also concerns polypeptides encoded by the BAP28 gene. The invention also deals with antibodies directed specifically against such polypeptides that are useful as diagnostic reagents. The invention further encompasses biallelic markers of the *BAP28* gene useful in genetic analysis, and more particularly associated with prostate cancer and useful in diagnosis.

## BACKGROUND OF THE INVENTION

### Prostate Cancer

The incidence of prostate cancer has dramatically increased over the last decades. It averages 30-50/100,000 males in Western European countries as well as within the US White male population. In these countries, it has recently become the most commonly diagnosed malignancy, being one of every four cancers diagnosed in American males. Prostate cancer's incidence is very much population specific, since it varies from 2/100,000 in China, to over 80/100,000 among African-American males.

In France, the incidence of prostate cancer is 35/100,000 males and it is increasing by 10/100,000 per decade. Mortality due to prostate cancer is also growing accordingly. It is the second cause of cancer death among French males, and the first one among French males aged over 70. This makes prostate cancer a serious burden in terms of public health.

Prostate cancer is a latent disease. Many men carry prostate cancer cells without overt signs of disease. Autopsies of individuals dying of other causes show prostate cancer cells in 30 % of men at age 50 and in 60 % of men at age 80. Furthermore, prostate cancer can take up to 10 years to kill a patient after the initial diagnosis.

The progression of the disease usually goes from a well-defined mass within the prostate to a breakdown and invasion of the lateral margins of the prostate, followed by metastasis to regional lymph nodes, and metastasis to the bone marrow. Cancer metastasis to bone is common and often associated with uncontrollable pain.

Unfortunately, in 80 % of cases, diagnosis of prostate cancer is established when the disease has already metastasized to the bones. Of special interest is the observation that prostate cancers frequently grow more rapidly in sites of metastasis than within the prostate itself.

Early-stage diagnosis of prostate cancer mainly relies today on Prostate Specific Antigen (PSA) dosage, and allows the detection of prostate cancer seven years before clinical symptoms become apparent. The effectiveness of PSA dosage diagnosis is however limited, due to its inability to discriminate between malignant and non-malignant affections of the organ and because not all prostate cancers give rise to an elevated serum PSA concentration. Furthermore, PSA dosage and other currently available approaches such as physical examination, tissue biopsy and bone scans are of limited value in predicting disease progression.

Therefore, there is a strong need for a reliable diagnostic procedure which would enable a more systematic early-stage prostate cancer prognosis.

Although an early-stage prostate cancer prognosis is important, the possibility of measuring the period of time during which treatment can be deferred is also interesting as currently available medicaments are expensive and generate important adverse effects. However, the aggressiveness of prostate tumors varies widely. Some tumors are relatively aggressive, doubling every six months whereas others are slow-growing, doubling once every five years. In fact, the majority of prostate cancers grows relatively slowly and never becomes clinically manifest. Very often, affected patients are among the elderly and die from another disease before prostate cancer actually develops. Thus, a significant question in treating prostate carcinoma is how to discriminate between tumors that will progress and those that will not progress during the expected lifetime of the patient.

Hence, there is also a strong need for detection means which may be used to evaluate the aggressiveness or the development potential of prostate cancer tumors once diagnosed.

Furthermore, at the present time, there is no means to predict prostate cancer susceptibility. It would also be very beneficial to detect individual susceptibility to prostate cancer. This could allow preventive treatment and a careful follow up of the development of the tumor.

A further consequence of the slow growth rate of prostate cancer is that few cancer cells are actively dividing at any one time, rendering prostate cancer generally resistant to radiation and chemotherapy. Surgery is the mainstay of treatment but it is largely ineffective and removes the ejaculatory ducts, resulting in impotence. Oral oestrogens and luteinizing releasing hormone analogs are also used for treatment of prostate cancer. These hormonal treatments provide marked improvement for many patients, but they only provide temporary relief. Indeed, most of these cancers soon relapse with the development of hormone-resistant tumor cells and the oestrogen treatment can lead to serious cardiovascular complications. Consequently, there is a strong need for preventive and curative treatment of prostate cancer.

Efficacy/tolerance prognosis could be precious in prostate cancer therapy. Indeed, hormonal therapy, the main treatment currently available, presents important side effects. The use of

chemotherapy is limited because of the small number of patients with chemosensitive tumors. Furthermore the age profile of the prostate cancer patient and intolerance to chemotherapy make the systematic use of this treatment very difficult.

Therefore, a valuable assessment of the eventual efficacy of a medicament to be administered to a prostate cancer patent as well as the patent's eventual tolerance to it may permit to enhance the benefit/risk ratio of prostate cancer treatment.

**BAP28**

Bowcock et al. (1998) conducted studies to identify proteins interacting with the the first 304 amino terminal amino acid residues of breast cancer related gene, BRCA1. Bowcock et al. thereby identified a BAP28 cDNA encoding a 515 amino acid protein associating with BRCA1 in a yeast two-hybrid screen, but whose association with BRCA1 could not be confirmed in a two-hybrid screen in mammalian cells.

## SUMMARY OF THE INVENTION

The present invention pertains to nucleic acid molecules comprising the genomic sequence of a novel human *BAP28* gene and BAP28 protein. The *BAP28* genomic sequence comprises regulatory sequences located upstream and downstream of the transcribed portion of said gene, these regulatory sequences being also part of the invention.

The invention also deals with complete cDNA sequences encoding the BAP28 protein, as well as with the corresponding translation product.

Oligonucleotide probes or primers hybridizing specifically with a *BAP28* genomic or cDNA sequence are also part of the present invention, as well as DNA amplification and detection methods using said primers and probes.

A further object of the invention consists of recombinant vectors comprising any of the nucleic acid sequences described herein, and in particular of recombinant vectors comprising a *BAP28* regulatory sequence or a sequence encoding a BAP28 protein, as well as of cell hosts and transgenic non human animals comprising said nucleic acid sequences or recombinant vectors.

The invention is also directed to BAP28 polymorphisms and BAP28-related biallelic markers as well as use of the of BAP28-related biallelic markers in establishing genetic associations with disease. BAP28-related biallelic markers can be used for diagnosis, staging, prognosis and monitoring of disease, and the efficient design and evaluation of suitable therapeutic solutions including individualized strategies for optimizing drug usage, and screening of potential new medicament candidates. More particularly, the invention concerns an association between BAP28-related biallelic markers and prostate cancer.

Finally, the invention is directed to methods for the screening of substances or molecules that inhibit the expression of *BAP28*, as well as with methods for the screening of substances or molecules that interact with a BAP28 polypeptide or that modulate the activity of a BAP28 polypeptide.

## BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a diagram showing the genomic structure of the genes BAP28 and PCTA-1. The arrow represent the DNA with the 5' to 3' direction. The boxes represent the exons.

Figure 2 is a diagram showing some alternative cDNA forms of the PCTA-1 gene.

Figure 3 is an alignment of the human BAP28 protein **H** with its homologues from *Drosophila melanogaster* (ORF from AE003615) **D**, *Arabidopsis thaliana* (AAF63640) **A**, *Schizosaccahromyces pombe* (O60179) **S**, *Caenorhabditis elegans* (Q23495) **C**, and *Saccharomyces cerevisiae* (YJK9_YEAST) **Y**. In C terminal part of the protein alignment, a box indicates the position of a conserved HEAT_REPEAT which is described to be involved in protein-protein interaction. For *Drosophila melanogaster*, the sequence AE003615 decribes a gene CG10805 with 6 exons. A new analysis showed that the exons 2, 3, 4, 5, and 6 present an holomoly with BAP28. Therefore, a new cDNA has been generated consisting with 21 bp upstream to exon 2, exon 2, intron 2, exons 3, 4, 5, and 6. This cDNA encodes a protein of 2096 amino acids which is described as D in the Figure 3.

Figure 4 is an alignment of the human BAP28 protein and 3 protein segments from *Tetraodon nigroviridis*, likely part of the same protein. The following sequences from Genbanl have been contigated in order to generate 3 segments of the genomic sequence of Tetraodon (CNS01RV3 + CNS03LT9 -->tetraodon3 ; CNS02AXF + CNS03INT --> tetraodon1 ; CNS02AXG + CNS01RV4 + CNS03LTA + CNS03INS --> tetraodon2). The 3 protein fragments which are similar to BAP28 have been found in these contigated regions. Furthermore, the exons encoding the 3 protein segments have the same size and the same structure in human BAP28 and in *Tetraodon*. The amino acid sequences encoding by these exons have been aligned with the human BAP28 protein.

Figure 5 is a diagram showing the allelic association analysis in chromosomic region 1q43.

Figure 6 is a diagram showing the genotypic association analysis in chromosomic region 1q43.

Figure 7 is a table demonstrating the results of a haplotype association analysis between prostate cancer cases and haplotypes comprising *BAP28*-related biallelic markers. Figure 7A a presents the results for the two-marker haplotypes. Figure 7B presents the results for the three-marker haplotypes.

Figure 8 is a table demonstrating the results of a haplotype association analysis between familial prostate cancer cases and haplotypes comprising *BAP28*-related biallelic markers. Figure 8A a presents the results for the two-marker haplotypes. Figure 8B presents the results for the three-marker haplotypes.

Figure 9 is a table demonstrating the results of a haplotype association analysis between early onset familial prostate cancer cases (less than 65 years old) and haplotypes comprising *BAP28*-related biallelic markers. Figure 9A a presents the results for the two-marker haplotypes. Figure 9B presents the results for the three-marker haplotypes.

Figure 10 is a table demonstrating the results of a haplotype association analysis between sporadic prostate cancer cases and haplotypes comprising *BAP28*-related biallelic markers. Figure 10A

a presents the results for the two-marker haplotypes. Figure 10B presents the results for the three-marker haplotypes.

Figure 11 is a table demonstrating the results of a haplotype association analysis between informative sporadic prostate cancer cases and haplotypes comprising *BAP28*-related biallelic markers. Figure 11A a presents the results for the two-marker haplotypes. Figure 11B presents the results for the three-marker haplotypes.

Figures12A and 12B are tables summarizing the results of haplotype frequency analyses between prostate cancer and three preferred haplotypes.

Figure 13 is a half-tome reproduction of the gels showing the tissular specificity of the BAP28 expression, more particularly the segment comprising the exons 43 to A. Figure 13 A : Wells 1 and 13: Molecular weight markers X - 300ng ; Well 2 : Mix PCR water = negative control ; Well 3 : Marathon Ready cDNA Human Testis : positive Tissue (CLONTECH Lot N°9110553) ; Well 4 : Marathon Ready cDNA Human Brain : negative Tissue ; Well 5 : Marathon Ready cDNA Human Cerebellum : negative Tissue ; Well 6 : Marathon Ready cDNA Human Cerebral Cortex : negative Tissue ; Well 7 : Marathon Ready cDNA Human Hippocampus : positive Tissue (CLONTECH Lot N°9040528) ; Well 8 : Marathon Ready cDNA Human Hypothalamus : negative Tissue ; Well 9 : Marathon Ready cDNA Human Fetal Kidney : negative Tissue ; Well 10 : Marathon Ready cDNA Human Thyroïd : negative Tissue ; Well 11 : Marathon Ready cDNA Human Bone Marrow : negative Tissue ; Well 11 : Marathon Ready cDNA Human Leukemia, promyelocytic HL60 : negative Tissue. Figure 13 B : Wells 1 and 7: Molecular weight markers X - 300ng ; Well 2 : Marathon Ready cDNA Human Leukemia, lymphoblastic MOLT4 : negative Tissue ;Well 3 : Marathon Ready cDNA Human Leukemia, chronic myelogenous K-562 : positive Tissue (CLONTECH Lot N°9120565) ; Well 4 : Marathon Ready cDNA Human Fetal Liver : negative Tissue ; Well 5 : Marathon Ready cDNA Human Stomach : negative Tissue ; Well 6 : Marathon Ready cDNA Human Prostate : negative Tissue. Figure 13C : Wells 1 and 13 : Molecular weight markers X - 300ng ; Well 2 : cDNA Human Testis : negative Tissue ; Well 3 : cDNA Human Cerebellum : positive Tissue (RNA PolyA+ CLONTECH -Lot N°8070047 - Réf Cat:6543-1) ; Well 4 : cDNA Human Corpus Callosum : negative Tissue ; Well 5 : cDNA Human Substantia Nigra : positive Tissue (RNA PolyA+ CLONTECH - Lot N°8090745 - Réf Cat:6580-1) ; Well 6 : cDNA Human Amygdala : negative Tissue ; Well 7 : cDNA Human Thalamus : positive Tissue (RNA PolyA+ CLONTECH -Lot N°9031131 - Réf Cat:6582-1) ; Well 8 : cDNA Human Hippocampus : positive Tissue (RNA PolyA+ CLONTECH -Lot N°8040059 - Réf Cat:6578-1) ; Well 9 : cDNA Human Caudate Nucleus : positive Tissue (RNA PolyA+ CLONTECH - Lot N°6120286 - Réf Cat:6575-1) ; Well 10 : cDNA Human Fetal Brain : negative Tissue ; Well 11 : cDNA Human Skeletal Muscle : negative Tissue ; Well 12 : cDNA Human Lung : negative Tissue. Figure 13 D : Wells 1 and 13 : Molecular weight markers X - 300ng ; Well 2 : cDNA Human Kidney : negative Tissue ; Well 3 : cDNA Human Placenta : negative Tissue ; Well 4 : cDNA Human Spleen : negative Tissue ; Well 5 : cDNA Human Fetal Liver : negative Tissue ; Well 6 :

cDNA Human Thyroïd Gland : negative Tissue ; Well 7 : cDNA Human Leukemia, lymphoblastic : negative Tissue ; Well 8 : cDNA Human Spinal Cord : positive Tissue (RNA PolyA+ CLONTECH - Lot N°9040709 - Réf Cat:6593-1) ; Well 9 : cDNA Human Pituitary Gland : positive Tissue (RNA PolyA+ CLONTECH - Lot N°6080167 - Réf Cat:6584-1) ; Well 10 : cDNA Human Adrenal Gland : negative Tissue ; Well 11 : cDNA Human Trachea : negative Tissue ; Well 12 : cDNA Human Leukemia, chronic myelogenous : negative Tissue. Figure 13 E : Wells 1 and 13 : Molecular weight markers X - 300ng ; Well 2 : cDNA Human Salivary Gland : negative Tissue ; Well 3 : cDNA Human Leukemia, promyelocytic : negative Tissue ; Well 4 : cDNA Human Small Intestine : negative Tissue ; Well 5 : cDNA Human Pancreas : negative Tissue ; Well 6 : cDNA Human Stomach : negative Tissue ; Well 7 : cDNA Human Mammary Gland : positive Tissue (RNA PolyA+ CLONTECH - Lot N°9031125 - Réf Cat:6545-1) ; Well 8 : cDNA Human Bone Marrow : negative Tissue ; Well 9 : cDNA Human Thymus : negative Tissue ; Well 10 : cDNA Human Uterus : negative Tissue ; Well 11 : cDNA Human Prostate : negative Tissue ; Well 12 : cDNA Human Prostate : negative Tissue.

Figure 14 is a block diagram of an exemplary computer system.

Figure 15 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database.

Figure 16 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 17 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence.

**Brief Description of the sequences provided in the Sequence Listing**

SEQ ID No 1 contains the genomic sequence of the *BAP28* gene comprising the exons and introns, and the 5' and 3' regulatory regions (respectively the upstream and downstream untranscribed regions). Furthermore, SEQ ID No 1 also contains the genomic sequence of the PCTA-1 gene. The coding strand of PCTA-1 gene is on the opposite of the coding strand of BAP28.

SEQ ID No 2 contains a first cDNA sequence of the *BAP28* gene consisting of the exons 1 to 45. SEQ ID No 3 contains a second cDNA sequence of the *BAP28* gene consisting of the exons 1 to 44, 45b and A'. SEQ ID No 4 contains a sequence of the *BAP28* cDNA segment consisting of the exons B' and A'. SEQ ID No 5 contains the BAP28 amino acid sequence encoded by the cDNAs of SEQ ID Nos 2, and 3.

SEQ ID No 6 contains a first cDNA sequence of the *PCTA-1* gene consisting of the exons 0 to 9. SEQ ID No 7 contains a second cDNA sequence of the *PCTA-1* gene consisting of the exons 0, 1, 2, 3, 4, 5, 6, 6bis, 7, 8, and 9. SEQ ID No 8 contains a third cDNA sequence of the *PCTA-1* gene consisting of the exons 0 to 8, 9bis and 9ter. SEQ ID No 9 contains the sequence of a cDNA fragment of the *PCTA-1* gene comprising exons C and A. SEQ ID No 10 contains the sequence of a cDNA

fragment of the *PCTA-1* gene comprising exons B, 0, 1 and 2. SEQ ID No 11 contains the sequence of a cDNA fragment of the *PCTA-1* gene comprising exons A, 1 and 2. SEQ ID No 12 contains the sequence of a cDNA fragment of the *PCTA-1* gene comprising exons A, D, 0, 1, and 2. SEQ ID No 13 contains a fourth cDNA sequence of the *PCTA-1* gene comprising exons A, 0, 1, 2, 3, 9bis and 9ter.

SEQ ID No 14 contains the PCTA-1 amino acid sequence encoded by the cDNAs of SEQ ID No 6. SEQ ID No 15 contains the PCTA-1 amino acid sequence encoded by the cDNAs of SEQ ID No 7. SEQ ID No 16 contains the PCTA-1 amino acid sequence encoded by the cDNAs of SEQ ID No 8. SEQ ID No 17 contains the PCTA-1 amino acid sequence encoded by the cDNAs of SEQ ID No 13.

SEQ ID Nos 18-31 contain the genomic amplicons respectively designated as 99-7177, 99-7212, 99-7193, 99-7186, 99-7182, 99-1585, 99-1587, 99-13798, 99-1601, 99-13808, 99-13810, 99-13790, 99-13809, and 99-1597.

SEQ ID Nos 31-61 contain the sequence of the following primers : BAP283Ra6283, BAP283Ra6324n, BAP28-exALF7311, BAP28-exALF7319n, PCTAexALF12, PCTAexALF13n, PCTAexALR60, PCTAexALR12n, PCTAexBLF33, PCTAexBLF120n, PCTAexBLR140, PCTAexBLR40n, PCTA5Ra220n, PCTA5Ra230, PCTA_5Ra400, PCTA_5Ran_400, PCTA_5Ra_394, PCTA_exD5Ra, PCTA_exD5Ran, PCTA_exC5Ra, PCTA_exC5Ran, PCTAex9terLR330, PCTAex9terLR325n, PCTAexCLF120, PCTAexCLF130n, BAP28polyTcourt, BAP281LF12.1, BAP28LR6726.1, BAP28LF26SalI and BAP28LR6717SalI, respectively.

SEQ ID No 62 contains a primer containing the additional PU 5' sequence described further in Example 2. SEQ ID No 63 contains a primer containing the additional RP 5' sequence described further in Example 2.

In accordance with the regulations relating to Sequence Listings, the following codes have been used in the Sequence Listing to indicate the locations of biallelic markers within the sequences and to identify each of the alleles present at the polymorphic base. The code "r" in the sequences indicates that one allele of the polymorphic base is a guanine, while the other allele is an adenine. The code "y" in the sequences indicates that one allele of the polymorphic base is a thymine, while the other allele is a cytosine. The code "m" in the sequences indicates that one allele of the polymorphic base is an adenine, while the other allele is an cytosine. The code "k" in the sequences indicates that one allele of the polymorphic base is a guanine, while the other allele is a thymine. The code "s" in the sequences indicates that one allele of the polymorphic base is a guanine, while the other allele is a cytosine. The code "w" in the sequences indicates that one allele of the polymorphic base is an adenine, while the other allele is an thymine. The nucleotide code of the original allele for each biallelic marker is the following:

| Biallelic marker | Original allele |
|---|---|
| A1 | G |
| A2 | C |
| A3 | T |
| A4 | C |
| A5 | C |
| A6 | T |
| A7 | T |
| A8 | G |
| A9 | T |

| A10 | G |
|---|---|
| A11 | G |
| A12 | A |
| A13 | T |
| A14 | T |
| A15 | A |
| A16 | G |
| A17 | T |
| A18 | T |
| A19 | C |
| A20 | G |
| **Biallelic marker** | **Original allele** |
| A21 | G |
| A22 | T |
| A23 | G |
| A24 | G |

| A25 | G |
|---|---|
| A26 | C |
| A27 | A |
| A28 | A |
| A29 | C |
| A30 | A |
| A31 | C |
| A32 | G |
| A33 | G |
| A34 | A |
| A35 | G |
| A36 | G |
| A37 | T |
| A38 | A |
| A39 | C |
| A40 | C |

In some instances, the polymorphic bases of the biallelic markers alter the identity of an amino acids in the encoded polypeptide. This is indicated in the accompanying Sequence Listing by use of the feature VARIANT, placement of an Xaa at the position of the polymorphic amino acid, and definition of Xaa as the two alternative amino acids. For example if one allele of a biallelic marker is the codon CAC, which encodes histidine, while the other allele of the biallelic marker is CAA, which encodes glutamine, the Sequence Listing for the encoded polypeptide will contain an Xaa at the location of the polymorphic amino acid. In this instance, Xaa would be defined as being histidine or glutamine.

In other instances, Xaa may indicate an amino acid whose identity is unknown because of nucleotide sequence ambiguity. In this instance, the feature UNSURE is used, placement of an Xaa at the position of the unknown amino acid and definition of Xaa as being any of the 20 amino acids or a limited number of amino acids suggested by the genetic code.

## DETAILED DESCRIPTION OF THE INVENTION

The present invention concerns polynucleotides and polypeptides related to the *BAP28* gene. Oligonucleotide probes and primers hybridizing specifically with a genomic or the cDNA sequences of *BAP28* are also part of the invention. A further object of the invention consists of recombinant vectors comprising any of the nucleic acid sequences described in the present invention, and in particular recombinant vectors comprising a regulatory region of *BAP28* or a sequence encoding the BAP28 protein, as well as cell hosts comprising said nucleic acid sequences or recombinant vectors. The invention also encompasses methods of screening of molecules which inhibit the expression of the *BAP28* gene or which modulate the activity of, or interact with, the BAP28 protein. The invention also deals with antibodies directed specifically against such polypeptides that are useful as diagnostic reagents.

The invention also concerns *BAP28*-related biallelic markers which can be used in any method of genetic analysis including linkage studies in families, linkage disequilibrium studies in populations and association studies of case-control populations. An important aspect of the present invention is that some BAP28-related biallelic markers present an association with the prostate cancer.

**Definitions**

Before describing the invention in greater detail, the following definitions are set forth to illustrate and define the meaning and scope of the terms used to describe the invention herein.

The terms "*BAP28* gene", when used herein, encompasses genomic, mRNA and cDNA sequences encoding the *BAP28* protein, including the untranslated regulatory regions of the genomic DNA.

The term "heterologous protein", when used herein, is intended to designate any protein or polypeptide other than the BAP28 protein. More particularly, the heterologous protein is a compound which can be used as a marker in further experiments with a *BAP28* regulatory region.

The term "isolated" requires that the material be removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition, and still be isolated in that the vector or composition is not part of its natural environment.

As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Purification of starting material or natural material is at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated. As an example, purification from 0.1 % concentration to 10 % concentration is two orders of magnitude.

To illustrate, individual cDNA clones isolated from a cDNA library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The cDNA clones are not naturally occurring as such, but rather are obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The conversion of mRNA into a cDNA library involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the synthetic library by clonal selection. Thus, creating a cDNA library from messenger RNA and subsequently isolating individual clones from that library results in an approximately $10^4$-$10^6$ fold purification of the native message.

The term "purified" is further used herein to describe a polypeptide or polynucleotide of the invention which has been separated from other compounds including, but not limited to,

polypeptides or polynucleotides, carbohydrates, lipids, etc. The term "purified" may be used to specify the separation of monomeric polypeptides of the invention from oligomeric forms such as homo- or hetero- dimers, trimers, etc. The term "purfied" may also be used to specify the separation of covalently closed polynucleotides from linear polynucleotides. A polynucleotide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polynucleotide sequence and conformation (linear versus covalently close). A substantially pure polypeptide or polynucleotide typically comprises about 50%, preferably 60 to 90% weight/weight of a polypeptide or polynucleotide sample, respectively, more usually about 95%, and preferably is over about 99% pure. Polypeptide and polynucleotide purity, or homogeneity, is indicated by a number of means well known in the art, such as agarose or polyacrylamide gel electrophoresis of a sample, followed by visualizing a single band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art. As an alternative embodiment, purification of the polypeptides and polynucleotides of the present invention may be expressed as "at least" a percent purity relative to heterologous polypeptides and polynucleotides (DNA, RNA or both). As a preferred embodiment, the polypeptides and polynucleotides of the present invention are at least: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 96%, 98%, 99%, or 100% pure relative to heterologous polypeptides and polynucleotides, respectively. As a further preferred embodiment the polypeptides and polynucleotides have a purity ranging from any number, to the thousandth position, between 90% and 100% (e.g., a polypeptide or polynucleotide at least 99.995% pure) relative to either heterologous polypeptides or polynucleotides, respectively, or as a weight/weight ratio relative to all compounds and molecules other than those existing in the carrier. Each number representing a percent purity, to the thousandth position, may be claimed as individual species of purity.

The term "polypeptide" refers to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude post-expression modifications of polypeptides, for example, polypeptides which include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

The term "recombinant polypeptide" is used herein to refer to polypeptides that have been artificially designed and which comprise at least two polypeptide sequences that are not found as contiguous polypeptide sequences in their initial natural environment, or to refer to polypeptides which have been expressed from a recombinant polynucleotide.

As used herein, the term "non-human animal" refers to any non-human vertebrate, birds and more usually mammals, preferably primates, farm animals such as swine, goats, sheep, donkeys, and horses, rabbits or rodents, more preferably rats or mice. As used herein, the term "animal" is used to refer to any vertebrate, preferable a mammal. Both the terms "animal" and "mammal" expressly embrace human subjects unless preceded with the term "non-human".

As used herein, the term "antibody" refers to a polypeptide or group of polypeptides which are comprised of at least one binding domain, where an antibody binding domain is formed from the folding of variable domains of an antibody molecule to form three-dimensional binding spaces with an internal surface shape and charge distribution complementary to the features of an antigenic determinant of an antigen, which allows an immunological reaction with the antigen. Antibodies include recombinant proteins comprising the binding domains, as wells as fragments, including Fab, Fab', $F(ab)_2$, and $F(ab')_2$ fragments.

As used herein, an "antigenic determinant" is the portion of an antigen molecule, in this case a BAP28 polypeptide, that determines the specificity of the antigen-antibody reaction. An "epitope" refers to an antigenic determinant of a polypeptide. An epitope can comprise as few as 3 amino acids in a spatial conformation which is unique to the epitope. Generally an epitope consists of at least 6 such amino acids, and more usually at least 8-10 such amino acids. Methods for determining the amino acids which make up an epitope include x-ray crystallography, 2-dimensional nuclear magnetic resonance, and epitope mapping e.g. the Pepscan method described by Geysen et al. 1984; PCT Publication No WO 84/03564; and PCT Publication No WO 84/03506.

Throughout the present specification, the expression "nucleotide sequence" may be employed to designate indifferently a polynucleotide or a nucleic acid. More precisely, the expression "nucleotide sequence" encompasses the nucleic material itself and is thus not restricted to the sequence information (i.e. the succession of letters chosen among the four base letters) that biochemically characterizes a specific DNA or RNA molecule.

As used interchangeably herein, the terms "nucleic acids", "oligonucleotides", and "polynucleotides" include RNA, DNA, or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form. The term "nucleotide" as used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. Although the term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar, for examples of analogous linking groups, purine, pyrimidines, and sugars see for example PCT

publication No WO 95/04064. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art.

A sequence which is "operably linked" to a regulatory sequence such as a promoter means that said regulatory element is in the correct location and orientation in relation to the nucleic acid to control RNA polymerase initiation and expression of the nucleic acid of interest.

As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence.

The terms "trait" and "phenotype" are used interchangeably herein and refer to any visible, detectable or otherwise measurable property of an organism such as symptoms of, or susceptibility to a disease for example. Typically the terms "trait" or "phenotype" are used herein to refer to symptoms of, or susceptibility to a disease, a beneficial response to or side effects related to a treatment. Preferably, said trait can be, without to be limited to, cancers, developmental diseases, and neurological diseases. More preferably, the term "trait" or "phenotype", when used herein, encompasses, but is not limited to, prostate cancer, an early onset of prostate cancer, a beneficial response to or side effects related to treatment or a vaccination against prostate cancer, a susceptibility to prostate cancer, the level of aggressiveness of prostate cancer tumors.

The term "allele" is used herein to refer to variants of a nucleotide sequence. A biallelic polymorphism has two forms. Typically the first identified allele is designated as the original allele whereas other alleles are designated as alternative alleles. The two alleles of a biallelic marker can also be referred to as allele 1 and allele 2. Diploid organisms may be homozygous or heterozygous for an allelic form.

The term "heterozygosity rate" is used herein to refer to the incidence of individuals in a population which are heterozygous at a particular allele. In a biallelic system, the heterozygosity rate is on average equal to $2P_a(1-P_a)$, where $P_a$ is the frequency of the least common allele. In order to be useful in genetic studies, a genetic marker should have an adequate level of heterozygosity to allow a reasonable probability that a randomly selected person will be heterozygous.

The term "genotype" as used herein refers the identity of the alleles present in an individual or a sample. In the context of the present invention, a genotype preferably refers to the description of the biallelic marker alleles present in an individual or a sample. The term "genotyping" a sample or an individual for a biallelic marker consists of determining the specific allele or the specific nucleotide carried by an individual at a biallelic marker.

The term "polymorphism" as used herein refers to the occurrence of two or more alternative genomic sequences or alleles between or among different genomes or individuals. "Polymorphic" refers to the condition in which two or more variants of a specific genomic sequence can be found in a population. A "polymorphic site" is the locus at which the variation occurs. A

single nucleotide polymorphism is the replacement of one nucleotide by another nucleotide at the polymorphic site. Deletion of a single nucleotide or insertion of a single nucleotide also gives rise to single nucleotide polymorphisms. In the context of the present invention, "single nucleotide polymorphism" preferably refers to a single nucleotide substitution.

The term "biallelic polymorphism" and "biallelic marker" are used interchangeably herein to refer to a single nucleotide polymorphism having two alleles at a fairly high frequency in the population. A "biallelic marker allele" refers to the nucleotide variants present at a biallelic marker site. Typically, the frequency of the less common allele of the biallelic markers of the present invention has been validated to be greater than 1%, preferably the frequency is greater than 10%, more preferably the frequency is at least 20% (i.e. heterozygosity rate of at least 0.32), even more preferably the frequency is at least 30% (i.e. heterozygosity rate of at least 0.42). A biallelic marker wherein the frequency of the less common allele is 30% or more is termed a "high quality biallelic marker".

The location of nucleotides in a polynucleotide with respect to the center of the polynucleotide are described herein in the following manner. When a polynucleotide has an odd number of nucleotides, the nucleotide at an equal distance from the 3' and 5' ends of the polynucleotide is considered to be "at the center" of the polynucleotide, and any nucleotide immediately adjacent to the nucleotide at the center, or the nucleotide at the center itself is considered to be "within 1 nucleotide of the center." With an odd number of nucleotides in a polynucleotide any of the five nucleotides positions in the middle of the polynucleotide would be considered to be within 2 nucleotides of the center, and so on. When a polynucleotide has an even number of nucleotides, there would be a bond and not a nucleotide at the center of the polynucleotide. Thus, either of the two central nucleotides would be considered to be "within 1 nucleotide of the center" and any of the four nucleotides in the middle of the polynucleotide would be considered to be "within 2 nucleotides of the center", and so on.

As used herein the term "BAP28-related biallelic marker" relates to a set of biallelic markers in linkage disequilibrium with the BAP28 gene or a BAP28 nucleotide sequence. The term "BAP28-related biallelic marker" relates to the biallelic markers located in a sequence selected from the group consisting of SEQ ID Nos 1-4, and 18-31, a fragment thereof and/or the complementary sequence thereto. The term *BAP28*-related biallelic marker encompasses the biallelic markers A1 to A58 disclosed in Table 2 and any biallelic markers in linkage disequilibrium therewith.

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym

from "complementary polynucleotide", "complementary nucleic acid" and "complementary nucleotide sequence". These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind.

**Variants and Fragments**

1- Polynucleotides

The invention also relates to variants and fragments of the polynucleotides described herein, particularly of a *BAP28* gene containing one or more biallelic markers according to the invention.

Variants of polynucleotides, as the term is used herein, are polynucleotides that differ from a reference polynucleotide. A variant of a polynucleotide may be a naturally occurring variant such as a naturally occurring allelic variant, or it may be a variant that is not known to occur naturally. Such non-naturally occurring variants of the polynucleotide may be made by mutagenesis techniques, including those applied to polynucleotides, cells or organisms. Generally, differences are limited so that the nucleotide sequences of the reference and the variant are closely similar overall and, in many regions, identical.

Variants of polynucleotides according to the invention include, without being limited to, nucleotide sequences which are at least 95% identical to a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 1-4, and 9-13 or to any polynucleotide fragment of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 consecutive nucleotides of a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 1-4 and 9-13, and preferably at least 99% identical, more particularly at least 99.5% identical, and most preferably at least 99.8% identical to a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 1-4 and 9-13, or to any polynucleotide fragment of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 consecutive nucleotides of a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID No 1-4 and 9-13.

Nucleotide changes present in a variant polynucleotide may be silent, which means that they do not alter the amino acids encoded by the polynucleotide. However, nucleotide changes may also result in amino acid substitutions, additions, deletions, fusions and truncations in the polypeptide encoded by the reference sequence. The substitutions, deletions or additions may involve one or more nucleotides. The variants may be altered in coding or non-coding regions or both. Alterations in the coding regions may produce conservative or non-conservative amino acid substitutions, deletions or additions.

In the context of the present invention, particularly preferred embodiments are those in which the polynucleotides encode polypeptides which retain substantially the same biological function or activity as the mature BAP28 protein, or those in which the polynucleotides encode

polypeptides which maintain or increase a particular biological activity, while reducing a second biological activity

A polynucleotide fragment is a polynucleotide having a sequence that is entirely the same as part but not all of a given nucleotide sequence, preferably the nucleotide sequence of a *BAP28* gene, and variants thereof. The fragment can be a portion of an intron or an exon of a *BAP28* gene. It can also be a portion of the regulatory regions of *BAP28*. In some embodiments, the fragments may comprise at least one polymorphism or biallelic marker of the invention.

Such fragments may be "free-standing", i.e. not part of or fused to other polynucleotides, or they may be comprised within a single larger polynucleotide of which they form a part or region. Indeed, several of these fragments may be present within a single larger polynucleotide.

In some embodiments, such fragments may comprise, consist of, or consist essentially of a contiguous span of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500 or 1000 nucleotides in length.

2- Polypeptides

The invention also relates to variants, fragments, analogs and derivatives of the polypeptides described herein, including mutated BAP28 proteins.

The variant may be 1) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue and such substituted amino acid residue may or may not be one encoded by the genetic code, or 2) one in which one or more of the amino acid residues includes a substituent group, or 3) one in which the mutated BAP28 is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or 4) one in which the additional amino acids are fused to the mutated BAP28, such as a leader or secretory sequence or a sequence which is employed for purification of the mutated BAP28 or a preprotein sequence. Such variants are deemed to be within the scope of those skilled in the art.

A polypeptide fragment is a polypeptide having a sequence that entirely is the same as part but not all of a given polypeptide sequence, preferably a polypeptide encoded by a *BAP28* gene and variants thereof.

In the case of an amino acid substitution in the amino acid sequence of a polypeptide according to the invention, one or several amino acids can be replaced by "equivalent" amino acids. The expression "equivalent" amino acid is used herein to designate any amino acid that may be substituted for one of the amino acids having similar properties, such that one skilled in the art of peptide chemistry would expect the secondary structure and hydropathic nature of the polypeptide to be substantially unchanged. Generally, the following groups of amino acids represent equivalent changes: (1) Ala, Pro, Gly, Glu, Asp, Gln, Asn, Ser, Thr; (2) Cys, Ser, Tyr, Thr; (3) Val, Ile, Leu, Met, Ala, Phe; (4) Lys, Arg, His; (5) Phe, Tyr, Trp, His.

A specific embodiment of a modified BAP28 peptide molecule of interest according to the present invention, includes, but is not limited to, a peptide molecule which is resistant to proteolysis, is a peptide in which the -CONH- peptide bond is modified and replaced by a (CH2NH) reduced bond, a (NHCO) retro inverso bond, a (CH2-O) methylene-oxy bond, a (CH2-S) thiomethylene bond, a (CH2CH2) carba bond, a (CO-CH2) cetomethylene bond, a (CHOH-CH2) hydroxyethylene bond), a (N-N) bound, a E-alcene bond or also a -CH=CH- bond. The invention also encompasses a human BAP28 polypeptide or a fragment or a variant thereof in which at least one peptide bond has been modified as described above.

Such fragments may be "free-standing", i.e. not part of or fused to other polypeptides, or they may be comprised within a single larger polypeptide of which they form a part or region. However, several fragments may be comprised within a single larger polypeptide.

As representative examples of polypeptide fragments of the invention, there may be mentioned those which have at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, 100 or 200 amino acids long. A specific embodiment of a BAP28 fragment is a fragment containing at least one amino acid mutation in the BAP28 protein.

**Identity Between Nucleic Acids Or Polypeptides**

The terms "percentage of sequence identity" and "percentage homology" are used interchangeably herein to refer to comparisons among polynucleotides and polypeptides, and are determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide or polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Homology is evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTP, FASTA, TFASTA, and CLUSTALW (Pearson and Lipman, 1988; Altschul et al., 1990; Thompson et al., 1994; Higgins et al., 1996; Altschul et al., 1990; Altschul et al., 1993). In a particularly preferred embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") which is well known in the art (see, e.g., Karlin and Altschul, 1990; Altschul et al., 1990, 1993, 1997). In particular, five specific BLAST programs are used to perform the following task: (1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database; (2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database; (3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database; (4) TBLASTN compares a

query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and, (5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (i.e., aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet et al., 1992; Henikoff and Henikoff, 1993). Less preferably, the PAM or PAM250 matrices may also be used (see, e.g., Schwartz and Dayhoff, eds., 1978). The BLAST programs evaluate the statistical significance of all high-scoring segment pairs identified, and preferably selects those segments which satisfy a user-specified threshold of significance, such as a user-specified percent homology. Preferably, the statistical significance of a high-scoring segment pair is evaluated using the statistical significance formula of Karlin (see, e.g., Karlin and Altschul, 1990).

**Stringent Hybridization Conditions**

For the purpose of defining such a hybridizing nucleic acid according to the invention, the stringent hybridization conditions are the followings :

the hybridization step is realized at 65°C in the presence of 6 x SSC buffer, 5 x Denhardt's solution, 0.5% SDS and 100µg/ml of salmon sperm DNA.

The hybridization step is followed by four washing steps :

- two washings during 5 min, preferably at 65°C in a 2 x SSC and 0.1% SDS buffer;
- one washing during 30 min, preferably at 65°C in a 2 x SSC and 0.1% SDS buffer,
- one washing during 10 min, preferably at 65°C in a 0.1 x SSC and 0.1%SDS buffer,

these hybridization conditions being suitable for a nucleic acid molecule of about 20 nucleotides in length. There is no need to say that the hybridization conditions described above are to be adapted according to the length of the desired nucleic acid, following techniques well known to the one skilled in the art. The suitable hybridization conditions may for example be adapted according to the teachings disclosed in the book of Hames and Higgins (1985).

**Table A**

| Exon | Position in SEQ ID No 1 | | Intron | Position in SEQ ID No 1 | |
|---|---|---|---|---|---|
| | Begining | End | | Begining | End |
| 1 | 4997 | 5076 | 1-2 | 5077 | 5370 |
| 2 | 5371 | 5544 | 2-3 | 5545 | 6120 |
| 3 | 6121 | 6337 | 3-4 | 6338 | 9876 |
| 4 | 9877 | 10018 | 4-5 | 10019 | 11521 |
| 5 | 11522 | 11623 | 5-6 | 11624 | 12520 |
| 6 | 12521 | 12661 | 6-7 | 12662 | 13452 |
| 7 | 13453 | 13664 | 7-8 | 13665 | 13823 |
| 8 | 13824 | 13957 | 8-9 | 13958 | 15375 |

| | | | | | |
|---|---|---|---|---|---|
| 9 | 15376 | 15478 | 9-10 | 15479 | 16854 |
| 10 | 16855 | 16965 | 10-11 | 16966 | 17377 |
| 11 | 17378 | 17495 | 11-12 | 17496 | 18534 |
| 12 | 18535 | 18642 | 12-13 | 18643 | 21445 |
| 13 | 21446 | 21541 | 13-14 | 21542 | 21998 |
| 14 | 21999 | 22087 | 14-15 | 22088 | 23035 |
| 15 | 23036 | 23247 | 15-16 | 23248 | 23545 |
| 16 | 23546 | 23667 | 16-17 | 23668 | 24269 |
| 17 | 24270 | 24461 | 17-18 | 24462 | 26286 |
| 18 | 26287 | 26470 | 18-19 | 26471 | 26610 |
| 19 | 26611 | 26747 | 19-20 | 26748 | 28067 |
| 20 | 28068 | 28260 | 20-21 | 28261 | 32539 |
| 21 | 32540 | 32709 | 21-22 | 32710 | 33111 |
| 22 | 33112 | 33270 | 22-23 | 33271 | 34585 |
| 23 | 34586 | 34828 | 23-24 | 34829 | 35155 |
| 24 | 35156 | 35287 | 24-25 | 35288 | 36659 |
| 25 | 36660 | 36763 | 25-26 | 36764 | 36933 |
| 26 | 36934 | 37077 | 26-27 | 37078 | 37802 |
| 27 | 37803 | 37921 | 27-28 | 37922 | 38016 |
| 28 | 38017 | 38138 | 28-29 | 38139 | 40364 |
| 29 | 40365 | 40493 | 29-30 | 40494 | 42617 |
| 30 | 42618 | 42848 | 30-31 | 42849 | 43451 |
| 31 | 43452 | 43578 | 31-32 | 43579 | 44835 |
| 32 | 44836 | 44999 | 32-33 | 45000 | 48222 |
| 33 | 48223 | 48269 | 33-34 | 48270 | 49655 |
| 34 | 49656 | 49779 | 34-35 | 49780 | 50357 |
| 35 | 50358 | 50498 | 35-36 | 50499 | 50963 |
| 36 | 50964 | 51256 | 36-37 | 51257 | 52147 |
| 37 | 52148 | 52298 | 37-38 | 52299 | 53234 |
| 38 | 53235 | 53393 | 38-39 | 53394 | 53553 |
| 39 | 53554 | 53688 | 39-40 | 53689 | 53837 |
| 40 | 53838 | 53942 | 40-41 | 53943 | 54028 |
| 41 | 54029 | 54197 | 41-42 | 54198 | 54740 |
| 42 | 54741 | 54895 | 42-43 | 54896 | 55753 |
| 43 | 55754 | 55912 | 43-44 | 55913 | 57385 |
| 44 | 57386 | 57494 | 44-45 | 57495 | 58503 |
| 45 | 58504 | 58827 | 45-B` | 58828 | 85946 |
| 45b | 58504 | 59354 | 45b-B` | 59355 | 85946 |
| B` | 85947 | 86168 | B`-A` | 86169 | 91228 |
| A` | 91229 | 91851 | | | |

**Genomic Sequences Of The Human *BAP28* Gene**

The present invention concerns the genomic sequence of *BAP28* comprising the sequence of SEQ ID No 1. The present invention encompasses *BAP28* gene, or *BAP28* genomic sequence consisting of, consisting essentially of, or comprising a sequence selected from the group consisting of SEQ ID No 1, a sequence complementary thereto, as well as fragments and variants thereof. These polynucleotides may be purified, isolated, or recombinant.

*BAP28* was localized by the present inventors to the chromosome 1q43 region.

The human *BAP28* genomic nucleic acid comprises at least 47 exons. The exon positions in SEQ ID No 1 are detailed below in the Table A.

The exons B' and A' of the *Bap28* gene have been found through the study of the *PCTA-1* gene which is described in the PCT application WO 99/64590, incorporated herein by reference. One public cDNA (Genbank Accession Number AF074001) shows an additional 5' exon in comparison of the cDNA described in the above-referenced application. This exon has been called exon B. It does not seem to comprise a splice site in 5'. So this exon will be a first exon. Long range PCR experiments with a first couple of primers PCTAexBLF33/PCTA5Ra230 (SEQ ID No 40/SEQ ID No 45) and a second one PCTAexBLF120n/PCTA5Ra220n (SEQ ID No 41/SEQ ID No 44) confirm the existence of a cDNA comprising at least the exon B and the exons 0, 1, and 2 (SEQ ID No 10).

Three additional exons have been also identified, namely exons A, C and D. Exon C is the most upstream exon. Exons A and D have a 5' splice site. Long range PCR with a first couple of primers PCTAexALF12/ PCTAex9terLR330 (SEQ ID No 36/SEQ ID No 53) and a second one PCTAexALF13n/ PCTAex9terLR325n (SEQ ID No 37/SEQ ID No 54) showed an alternative PCTA-1 cDNA consisting with the exons A, 0, 1, 2, 3, 9bis and 9ter (SEQ ID No 13). Other alternative PCTA-1 cDNAs comprise consecutively the exons A, D, 0, 1, and 2 (SEQ ID No 12), the exons A, 1 and 2 (SEQ ID No 11), or the exons C and A (SEQ ID No 9). The form AD012 and A12 have been amplified with the first couple of primers PCTAexALF12 / PCTA5Ra230 (SEQ ID No 36/SEQ ID No 45) and the second one PCTAexALF13n /PCTA5Ra220n (SEQ ID No 37/SEQ ID No 44). The exon C have been identified by a RACE experiment with PCTAexALR60 primer (SEQ ID No 38) from the exon A. The figure 2 shows the alternative cDNAs of *PCTA-1* and the alternative 5' ends of *PCTA-1* cDNAs.

The first identified *BAP28* cDNAs comprise either the exons 1 to 45 or 1 to 44 and 45b. They are detailed in the section "*BAP28* cDNA sequences". The exon 45 of the *BAP28* cDNA comprises a polyadenylation site and some RACE experiments failed not show any additional sequence downstream of the exon 45, which was the last identified exon.

The study of the PCTA-1 new exons for an alternative cDNA comprising both the exons A and B provides two additional *BAP28* exons, the exons A' and B'. Indeed, two upstream PCR primers were designed; one in the exon A (PCTAexALF12 (SEQ ID No 36 following by PCTAexALF13n (SEQ ID No 37)) and the other in exon B (PCTAexBLF33 (SEQ ID No 40) following by PCTAexBLF120n (SEQ ID No 41)). The downstream primer was generated in previously identified PCTA-1 exons (PCTA5Ra230 (SEQ ID No 45) following by PCTA5Ra220n (SEQ ID No 44)). No alternative cDNA comprising both exons has been observed. Therefore, two couples of primers was designed with the upstream primer in exon A and the downstream primer in exon B. More particularly, the amplification was done with a first couple of primers PCTAexALF12/ PCTAexBLR140 (SEQ ID No 36/SEQ ID No 42) and a second one

PCTAexALF13n/ PCTAexBLR40n (SEQ ID No 37/SEQ ID No 43). An amplification product was obtained. However, the exons were slightly moved and the splice sites were only available on the opposite strand. Therefore, the amplification product was not from the *PCTA-1* gene but rather than was supposed to be from the *BAP28* gene which is on the opposite strand. This amplification product contains the exons A` and B` (SEQ ID No 4). In order to check that the amplification product comes from BAP28, a PCR amplification was proceeded with a dowstream primer in the exon A and an upstream primer in exon 43 of BAP28 gene. More particularly, the PCR was done with a first couple of primers PCTAexALF12/ BAP283Ra6283 (SEQ ID No 36/SEQ ID No 32) and a second one PCTAexALF13n/ BAP283Ra6324n (SEQ ID No 37/SEQ ID No 33) The amplification product confirmed that the slightly moved exons A and B are part of the *BAP28* cDNA. The sequencing of the amplification product showed a cDNA comprising the exons 44, 45b, and A. The *BAP28* cDNA with the exons B` and A` likely consists to an other alternative cDNA form.

Thus, the invention embodies purified, isolated, or recombinant polynucleotides comprising a nucleotide sequence selected from the group consisting of the exons of the *BAP28* gene, or a sequence complementary thereto. Preferred are nucleotide sequences selected from the group consisting of the exons of the *BAP28* gene having the nucleotide position ranges listed in Table A, or a complementary sequence thereto or a fragment or a variant thereof.

Encompassed by the invention are purified, isolated, or recombinant nucleic acids comprising a combination of at least two exons of the *BAP28* gene, wherein the polynucleotides are arranged within the nucleic acid, from the 5`-end to the 3`-end of said nucleic acid, in the same order as in SEQ ID No 1. The invention further deals with purified, isolated, or recombinant nucleic acids comprising a combination of at least two exons of the *BAP28* gene, wherein the nucleic acids comprise at least one exon selected from the group consisting of exons 1 to 45, 45b, B` and A`, wherein the polynucleotides are arranged within the nucleic acid, from the 5`-end to the 3`-end of said nucleic acid, in the same order as in SEQ ID No 1.

Preferred polynucleotides of the invention embody purified, isolated, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, or 200 nucleotides, to the extent that such a length is consistent with the lengths of the particular nucleotide position, of SEQ ID No 1 or the complement thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 20, 30, 40 or 50 nucleotides selected from the group consisting of the following nucleotide positions of SEQ ID No 1: 4997-5076, 5371-5544, 6121-6337, 9877-10018, 11522-11623, 12521-12661, 13453-13664, 13824-13957, 15376-15478, 16855-16965, 17378-17495, 18535-18642, 21446-21541, 21999-22087, 23036-23247, 23546-23667, 24270-24461, 26287-26470, 26611-26747, 28068-28260, 32540-32709, 33112-33270, 34586-34828, 35156-35287, 36660-36763, 36934-37077, 37803-37921, 38017-38138, 40365-40493, 42618-42848, 43452-43578, 44836-44999, 48223-48269, and 49656-49779.

The position of the introns is detailed in Table A. Thus, the invention embodies purified, isolated, or recombinant polynucleotides comprising a nucleotide sequence selected from the group consisting of the introns of the *BAP28* gene, or a sequence complementary thereto.

The invention also encompasses a purified, isolated, or recombinant polynucleotides comprising a nucleotide sequence having at least 70, 75, 80, 85, 90, or 95% nucleotide identity with a nucleotide sequence of SEQ ID No 1 or a complementary sequence thereto or a fragment thereof. The nucleotide differences as regards to the nucleotide sequences of SEQ ID No 1 may be generally randomly distributed throughout the entire nucleic acid. Nevertheless, preferred nucleic acids are those wherein the nucleotide differences as regards to the nucleotide sequences of SEQ ID No 1 are predominantly located outside the coding sequences contained in the exons. These nucleic acids, as well as their fragments and variants, may be used as oligonucleotide primers or probes in order to detect the presence of a copy of the *BAP28* gene in a test sample, or alternatively in order to amplify a target nucleotide sequence within the *BAP28* sequences.

Another object of the invention consists of a purified, isolated, or recombinant nucleic acids that hybridizes with a nucleotide sequence selected from the group consisting of SEQ ID No 1 or a complementary sequence thereto or a variant thereof, under the stringent hybridization conditions as defined above.

Particularly preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides, to the extent that such a length is consistent with the lengths of the particular nucleotide position, of SEQ ID No 1 or the complement thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 20, 30, 40 or 50 of the following nucleotide positions of SEQ ID No 1: 1-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662.

Further preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides, to the extent that such a length is consistent with the lengths of the particular nucleotide position, of SEQ ID No 1 or the complement thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 20, 30, 40 or 50 of the following nucleotide positions of SEQ ID No 1: 1-2500, 2501-5000, 5001-7500, 7501-10000, 10001-12500, 12501-15000, 15001-17500, 17501-20000, 20001-22500, 22501-25000, 25001-27500, 27501-30000, 30001-32500, 32501-35000, 35001-37500, 37501-40000, 40001-42500, 42501-45000, 45001-47500, 47501-50000, 50001-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662.

Other preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides, to the extent that such a length is consistent with the lengths of the particular nucleotide position, of SEQ ID No 1, or the complements thereof, wherein said contiguous span comprises at least one *BAP28*-related biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably at least one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53.

It should be noted that nucleic acid fragments of any size and sequence may also be comprised by the polynucleotides described in this section.

In another aspect, the invention concerns polymorphisms of *BAP28*.

While this section is entitled "Genomic Sequences of *BAP28*," it should be noted that nucleic acid fragments of any size and sequence may also be comprised by the polynucleotides described in this section, flanking the genomic sequences of *BAP28* on either side or between two or more such genomic sequences.

### *BAP28* cDNA Sequences

Another object of the invention is a purified, isolated, or recombinant nucleic acid comprising a nucleotide sequence selected from the group consisting of SEQ ID Nos 2 and 3, complementary sequences thereto, as well as allelic variants, and fragments thereof. Moreover, preferred polynucleotides of the invention include purified, isolated, or recombinant *BAP28* cDNAs consisting of, consisting essentially of, or comprising a nucleotide sequence selected from the group consisting of SEQ ID Nos 2 and 3. The two *BAP28* cDNAs have to a different 3' end. The first one, namely the cDNA of the SEQ ID No 2, comprises the exons 1 to 44 and 45. The second one, namely the cDNA of the SEQ ID No 3, comprises the exons 1 to 44, 45b and A'. The cDNA of SEQ ID No 2 or 3 are described in Table B.

Consequently, the invention concerns a purified, isolated, and recombinant nucleic acids comprising a nucleotide sequence of the 5'UTR of the *BAP28* cDNA, a sequence complementary thereto, or an allelic variant thereof. The invention also concerns a purified, isolated, and recombinant nucleic acids comprising a nucleotide sequence of the 3'UTR of the *BAP28* cDNA, a sequence complementary thereto, or an allelic variant thereof.

**Table B**

| cDNA | Position range of 5UTR | | Position range of ORF | | Position range of 3UTR | |
|---|---|---|---|---|---|---|
| **cDNA1** | 1 | 112 | 113 | 6547 | 6548 | 6782 |
| **cDNA2** | 1 | 112 | 113 | 6547 | 6548 | 7932 |

As described in the section "Genomic Sequences of the human *Bap28* gene", an alternative form of the *BAP28* cDNA comprises the exons B' and A'. Therefore, the invention concerns a cDNA of BAP28 comprising the nucleotide sequence of SEQ ID No 4.

Particularly preferred embodiments of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of nucleotide positions 1 to 4995 of SEQ ID No 2 or 3. Further preferred polynucleotides include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 2 or 3: 1 to 2033, 2160 to 2348, and 2676 to 4995. Additional preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of SEQ ID No 2, or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the following ranges of nucleotide positions of SEQ ID No 2: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, and 6501 to 6782. Additional preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of SEQ ID No 3, or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the following ranges of nucleotide positions of SEQ ID No 3: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, 6501 to 7000, 7001 to 7500, 7501 to 7932.

The invention also pertains to a purified or isolated nucleic acid having at least 95% of nucleotide identity with a nucleotide sequence selected from the group consisting of SEQ ID Nos 2 and 3 or a fragment thereof or a complementary sequence thereto, advantageously 99 %, preferably 99.5% nucleotide identity and most preferably 99.8% nucleotide identity with a nucleotide sequence selected from the group consisting of SEQ ID Nos 2 and 3 or a fragment thereof or a complementary sequence thereto.

Another object of the invention consists of a purified, isolated, or recombinant nucleic acids that hybridizes with a nucleotide sequence selected from the group consisting of SEQ ID Nos 2 and 3 or a complementary sequence thereto or a variant thereof, under the stringent hybridization conditions as defined above.

The invention concerns a PCTA-1 cDNA comprising an exon selected from the group consisting of exons A, B, C, and D. More particularly, the invention concerns a PCTA-1 cDNA comprising a polynucleotide sequence selected from the group consisting of SEQ ID Nos 9-13 or a fragment thereof or a complementary sequence thereto.

Encompassed by the invention are purified, isolated, or recombinant nucleic acids comprising a combination of at least two exons of the *PCTA-1* gene, wherein the polynucleotides are arranged within the nucleic acid, from the 5`-end to the 3`-end of said nucleic acid, in the same order as in SEQ ID No 1. The invention further deals with purified, isolated, or recombinant nucleic acids comprising a combination of at least two exons of the *PCTA-1* gene, wherein the nucleic acids comprise at least one exon selected from the group consisting of exons C, A, D, B, 0, 1, 2, 3, 4, 5, 6, 6bis, 7, 8, 9, 9bis and 9ter, wherein the polynucleotides are arranged within the nucleic acid, from the 5`-end to the 3`-end of said nucleic acid, in the same order as in SEQ ID No 1.

While this section is entitled "*BAP28* cDNA Sequences," it should be noted that nucleic acid fragments of any size and sequence may also be comprised by the polynucleotides described in this section, flanking the genomic sequences of *BAP28* on either side or between two or more such genomic sequences.

## NATURAL ANTISENSE

Over the last 10 years, an increasing number of natural antisense RNAs has been reported in eukaryotes. Natural antisense RNAs are endogenous transcripts that exhibit complementary sequences to other transcripts, named sense transcripts. Most antisense transcripts are issued from the same locus as sense transcripts. Transcribed from opposite strands of DNA, sense and antisense transcripts overlap each other at least partially, and display perfect complementarity. The reported antisense RNAs are complementary to sense transcripts encoding proteins involved in extremely diverse biological functions : hormonal response, control of proliferation, development, structure, etc...

In some cases, apart from their capability of encoding proteins *per se*, antisense RNAs were found to regulate, generally downregulate, the expression of their sense counterparts. Often changes in sense gene expression were correlated with the presence of antisense RNA. Indeed, an inverse relationship between levels of accumulation of sense and antisense messengers has been documented in several cases. Some examples have been reported in various pathology such as nervous disorders and cancer.

These characteristics suggest that antisense transcripts are found throughout the whole eukaryotic world and might play a role in general antisense-mediated gene regulation as is the cases in prokaryotes. Indeed, antisense –mediated gene regulation is a way of decreasing the abundance of stable transcripts more rapidly than the cessation of transcription. In addition, natural antisense

transcripts are thought to be involved not only in the normal regulation of gene expression but also in the alteration of gene regulation leading to different pathologies.

Indeed, because of their complementarity, antisense transcripts may hybridize to sense transcripts and thus modify the expression of their sense counterparts at any step from transcription to translation.

In the nucleus, antisense RNA may regulate sense expression either at the level of transcription, processing, or nucleocytoplasmic transport. Transcriptional regulation occurs either because the activity of sense and antisense promoters is differentially regulated by cellular conditions or because antisense transcription impedes sense transcription. This interference would involve the collision of two transcription complexes, resulting in premature termination or in reduced elongation of transcription, the transcripts with the highest rate of transcription being predominant. Antisense may also operate at a post-transcriptional level probably by impairing either maturation and/or transport of the sense transcript.

Although some examples have shown that antisense regulation may occur in the nucleus, antisense regulation is generally described as a cytoplasmic event operating mostly at the messenger stability level. Furthermore, the regulation can also be made at the translation stage, particularly when interactions between sense and antisense occur in the 3'UTR.

Two mechanisms of antisense-mediated gene regulation may be envisioned. First, antisense transcripts displaying very similar structural features to sense transcripts may bind proteins actually interacting with their sense counterparts, thus depriving sense messengers from proteins necessary for their functions. The other mechanism of antisense-mediated regulation is thought to operate via duplex formation between complementary sense and antisense transcripts. By simple steric hindrance, RNA duplexes would prevent sense RNA from interacting with diverse cellular components required for normal sense expression, thus impairing maturation, nucleocytoplasmic transport, transcript stability, or translation depending on the cellular components involved. Alternatively, duplexes may represent substrates for double-stranded RNA specific enzymes. It is commonly believed that most duplexes will become targeted for degradation by RNAses and only the most abundant transcripts, either sense or antisense, will persist in the cells. More information on the natural antisense can be found in Vanhee-Brossollet et al. (1998).

**BAP28 and PCTA-1 are natural antisense**

*BAP28* transcript has been identified as a natural antisense of the *PCTA-1* transcript. Indeed, the coding sequence of *PCTA-1* is on the opposite strand of the coding sequence of *BAP28*. Moreover, the 3'UTR of *BAP28* contains some sequences which are complementary of segments of the 5'UTR and 3'UTR of *PCTA-1*. More particularly, the exons A and B are common for the *PCTA-1* and *BAP28* genes, the exon 44 of *BAP28* gene is antisense of the exons 9 and 9ter of *PCTA-1*, the exons 45 and 45b of *BAP28* gene are antisense of the exon 9 of *PCTA-1*. Therefore, *BAP28*

transcript is the antisense of the *PCTA-1* RNA. The Figure 1 presents the general organization of the *BAP28* and *PCTA-1* genes.

The PCTA-1 protein has been shown to be a specific antigen of prostate cancer cells (WO 96/21671, incorporated herein by reference). Therefore, one can assume that its expression is closely linked to the development of cancer, particularly prostate cancer.

ESTs from the *PCTA-1* gene were found in a broad range of tissues. As the protein PCTA-1 is only present in the prostate cancer cells, a regulation of the *PCTA-1* RNA will occur, maybe at the stage of the RNA transcription, splicing, stability and/or translation.

The 5'UTR and 3'UTR regions of a gene are of particular importance in that they often comprise regulatory elements which can play a role in providing appropriate expression levels, particularly through the control of mRNA stability.

As the *BAP28* transcript is the natural antisense of the *PCTA-1* mRNA, the *BAP28* mRNA is likely to be involved in the regulation of the *PCTA-1* expression and, by consequence, in the process of development of prostate cancer.

The involvement of *BAP28* gene in prostate cancer is reported through the clearly significant association of the *BAP28*-related biallelic markers to prostate cancer. Furthermore, the PCT application WO98/12327, incorporated herein by reference, showed that BAP28 should be involved in interaction with BRCA1. Therefore, BAP28 may be a tumor suppressor. During the process of carcinogenesis, BAP28 would become inactive and its expression could decrease. This expression decrease of BAP28 would lead to an increase of the PCTA-1 mRNA stability and the presence of the PCTA-1 protein at the cell surface. We can hypothesize that these events correspond to a natural defense against the cancer cells.

Consequently, the invention concerns the use of BAP28 nucleotide sequence from the mRNA as antisense in order to control the PCTA-1 expression and preferably to inhibit the PCTA-1 expression. The invention also concerns the use of PCTA-1 nucleotide sequence from the mRNA as an antisense in order to control the BAP28 expression. These antisense can be used in order to avoid cancer development, preferably prostate cancer development.

An embodiment of the invention concerns the polynucleotide segment common in the PCTA-1 and BAP28 cDNAs. More particularly, the invention concerns isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of SEQ ID No 1, or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the following ranges of nucleotide positions of SEQ ID No 1: 57386-27494, 58504-59354, 85947-86108, and 91259-91325.

An additional embodiment is the use of a polynucleotide according to the invention, more particularly polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of SEQ ID No 1, or the

complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the following ranges of nucleotide positions of SEQ ID No 1: 57386-27494, 58504-59354, 85947-86108, and 91259-91325, for regulating the expression of PCTA-1 and/or BAP28.

**Coding Regions**

The *BAP28* open reading frame is contained in the corresponding mRNAS of SEQ ID No 2 or 3. More precisely, the effective *BAP28* coding sequence (CDS) includes the region between nucleotide position 113 (first nucleotide of the ATG codon) and nucleotide position 6547 (end nucleotide of the TAA codon) of SEQ ID No 2 or 3.

Thus, the present invention deals with a purified or isolated nucleic acid encoding a BAP28 protein or a fragment thereof. More particularly the present invention deals with a purified or isolated nucleic acid encoding a BAP28 protein having the amino acid sequence of SEQ ID No 5 or a peptide fragment or variant thereof. The present invention also embodies isolated, purified, and recombinant polynucleotides which encode a polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 or 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes at least 1, 2, 3, 5 or 10 of the amino acid positions 1 to 1629 of the SEQ ID No 5. The present invention further embodies isolated, purified, and recombinant polynucleotides which encode a polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 or 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span contains an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5 and an alanine at the amino acid position 2050 of SEQ ID No 5. The present invention embodies isolated, purified, and recombinant polynucleotides which encode a polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 or 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes at least 1, 2, 3, 5 or 10 of the amino acid positions 1 to 200, 201 to 400, 401 to 600, 601 to 800, 801 to 1000, 1001 to 1200, 1201 to 1400 and/or 1401 to 1629 of the SEQ ID No 5.

The above disclosed polynucleotide that contains the coding sequence of the *BAP28* gene may be expressed in a desired host cell or a desired host organism, when this polynucleotide is placed under the control of suitable expression signals. The expression signals may be either the expression signals contained in the regulatory regions in the *BAP28* gene of the invention or in contrast the signals may be exogenous regulatory nucleic sequences. Such a polynucleotide, when placed under the suitable expression signals, may also be inserted in a vector for its expression and/or amplification.

**Regulatory Sequences Of *BAP28***

As mentioned, the genomic sequence of the *BAP28* gene contains regulatory sequences both in the non-coding 5`-flanking region and in the non-coding 3`-flanking region that border the *BAP28* coding region containing the 45 exons of this gene.

The 5`-regulatory sequence of the *BAP28* gene is localized between the nucleotide in position 2996 and the nucleotide in position 4996 of the nucleotide sequence of SEQ ID No 1. The 5`-regulatory sequence contains the *BAP28* promoter site.

The genomic sequence of the *BAP28* gene also contains regulatory sequences in the non-coding 3`-flanking region that border the *BAP28* coding region. The 3`-regulatory sequence of the *BAP28* gene is localized between nucleotide position 91852 and nucleotide position 97662 of SEQ ID No 1.

Polynucleotides derived from the 5` and 3` regulatory regions are useful in order to detect the presence of at least a copy of a nucleotide sequence of SEQ ID No 1 or a fragment thereof in a test sample.

The promoter activity of the 5` regulatory regions contained in *BAP28* can be assessed as described below.

In order to identify the relevant biologically active polynucleotide fragments or variants of SEQ ID No 1, the one skill in the art will refer to the book of Sambrook et al.(1989) which describes the use of a recombinant vector carrying a marker gene (i.e. beta galactosidase, chloramphenicol acetyl transferase, etc.) the expression of which will be detected when placed under the control of a biologically active polynucleotide fragments or variants of SEQ ID No 1. Genomic sequences located upstream of the first exon of the *BAP28* gene are cloned into a suitable promoter reporter vector, such as the pSEAP-Basic, pSEAP-Enhancer, pβgal-Basic, pβgal-Enhancer, or pEGFP-1 Promoter Reporter vectors available from Clontech, or pGL2-basic or pGL3-basic promoterless luciferase reporter gene vector from Promega. Briefly, each of these promoter reporter vectors include multiple cloning sites positioned upstream of a reporter gene encoding a readily assayable protein such as secreted alkaline phosphatase, luciferase, β galactosidase, or green fluorescent protein. The sequences upstream the *BAP28* coding region are inserted into the cloning sites upstream of the reporter gene in both orientations and introduced into an appropriate host cell. The level of reporter protein is assayed and compared to the level obtained from a vector which lacks an insert in the cloning site. The presence of an elevated expression level in the vector containing the insert with respect to the control vector indicates the presence of a promoter in the insert. If necessary, the upstream sequences can be cloned into vectors which contain an enhancer for increasing transcription levels from weak promoter sequences. A significant level of expression above that observed with the vector lacking an insert indicates that a promoter sequence is present in the inserted upstream sequence.

Promoter sequence within the upstream genomic DNA may be further defined by constructing nested 5' and/or 3' deletions in the upstream DNA using conventional techniques such as Exonuclease III or appropriate restriction endonuclease digestion. The resulting deletion fragments can be inserted into the promoter reporter vector to determine whether the deletion has reduced or obliterated promoter activity, such as described, for example, by Coles et al.(1998). In this way, the boundaries of the promoters may be defined. If desired, potential individual regulatory sites within the promoter may be identified using site directed mutagenesis or linker scanning to obliterate potential transcription factor binding sites within the promoter individually or in combination. The effects of these mutations on transcription levels may be determined by inserting the mutations into cloning sites in promoter reporter vectors. This type of assay is well-known to those skilled in the art and is described in WO 97/17359, US 5,374,544; EP 582 796; US 5,698,389; US 5,643,746; US 5,502,176; and US 5,266,488; incorporated herein by reference.

The strength and the specificity of the promoter of the *BAP28* gene can be assessed through the expression levels of a detectable polynucleotide operably linked to the *BAP28* promoter in different types of cells and tissues. The detectable polynucleotide may be either a polynucleotide that specifically hybridizes with a predefined oligonucleotide probe, or a polynucleotide encoding a detectable protein, including a *BAP28* polypeptide or a fragment or a variant thereof. This type of assay is well-known to those skilled in the art and is described in US 5,502,176; and US 5,266,488; incorporated herein by reference. Some of the methods are discussed in more detail below.

Polynucleotides carrying the regulatory elements located at the 5' end and at the 3' end of the *BAP28* coding region may be advantageously used to control the transcriptional and translational activity of heterologous polynucleotide of interest.

Thus, the present invention also concerns a purified or isolated nucleic acid comprising a polynucleotide which is selected from the group consisting of the 5' and 3' regulatory regions, or a sequence complementary thereto or a biologically active fragment or variant thereof.

The invention also pertains to a purified or isolated nucleic acid comprising a polynucleotide having at least 95% nucleotide identity with a polynucleotide selected from the group consisting of the 5' and 3' regulatory regions, advantageously 99 % nucleotide identity, preferably 99.5% nucleotide identity and most preferably 99.8% nucleotide identity with a polynucleotide selected from the group consisting of the 5' and 3' regulatory regions, or a sequence complementary thereto or a variant thereof or a biologically active fragment thereof.

Another object of the invention consists of purified, isolated or recombinant nucleic acids comprising a polynucleotide that hybridizes, under the stringent hybridization conditions defined herein, with a polynucleotide selected from the group consisting of the nucleotide sequences of the 5'- and 3' regulatory regions, or a sequence complementary thereto or a variant thereof or a biologically active fragment thereof.

Preferred fragments of either the 5' or 3' regulatory region have a length of about 1500 or 1000 nucleotides, preferably of about 500 nucleotides, more preferably about 400 nucleotides, even more preferably 300 nucleotides and most preferably about 200 nucleotides.

By "biologically active" polynucleotide derivatives of SEQ ID No 1 are polynucleotides comprising or alternatively consisting in a fragment of said polynucleotide which is functional as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide in a recombinant cell host. It could act either as an enhancer or as a repressor.

For the purpose of the invention, a nucleic acid or polynucleotide is "functional" as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide if said regulatory polynucleotide contains nucleotide sequences which contain transcriptional and translational regulatory information, and such sequences are "operably linked" to nucleotide sequences which encode the desired polypeptide or the desired polynucleotide.

The regulatory polynucleotides of the invention may be prepared from the nucleotide sequence of SEQ ID No 1 by cleavage using suitable restriction enzymes, as described for example in the book of Sambrook et al.(1989). The regulatory polynucleotides may also be prepared by digestion of SEQ ID No 1 by an exonuclease enzyme, such as Bal31 (Wabiko et al., 1986). These regulatory polynucleotides can also be prepared by nucleic acid chemical synthesis, as described elsewhere in the specification.

The regulatory polynucleotides according to the invention may be part of a recombinant expression vector that may be used to express a coding sequence in a desired host cell or host organism. The recombinant expression vectors according to the invention are described elsewhere in the specification.

A preferred 5'-regulatory polynucleotide of the invention thus includes the 5'-UTR of the *BAP28* cDNA, or a biologically active fragment or variant thereof.

A preferred 3'-regulatory polynucleotide of the invention includes the 3'-UTR of the *BAP28* cDNA, or a biologically active fragment or variant thereof.

A further object of the invention consists of a purified or isolated nucleic acid comprising:

a) a nucleic acid comprising a regulatory nucleotide sequence selected from the group consisting of:

(i) a nucleotide sequence comprising a polynucleotide of the 5' regulatory region or a complementary sequence thereto;

(ii) a nucleotide sequence comprising a polynucleotide having at least 95% of nucleotide identity with the nucleotide sequence of the 5' regulatory region or a complementary sequence thereto;

(iii) a nucleotide sequence comprising a polynucleotide that hybridizes under stringent hybridization conditions with the nucleotide sequence of the 5' regulatory region or a complementary sequence thereto; and

(iv) a biologically active fragment or variant of the polynucleotides in (i), (ii) and (iii);

b) a polynucleotide encoding a desired polypeptide or a nucleic acid of interest, operably linked to the nucleic acid defined in (a) above;

c) In some embodiments, a nucleic acid comprising a 3`- regulatory polynucleotide, preferably a 3`- regulatory polynucleotide of the *BAP28* gene.

In a specific embodiment of the nucleic acid defined above, said nucleic acid includes the 5`-UTR of the *BAP28* cDNA, or a biologically active fragment or variant thereof.

In a second specific embodiment of the nucleic acid defined above, said nucleic acid includes the 3`-UTR of the *BAP28* cDNA, or a biologically active fragment or variant thereof.

The desired polypeptide encoded by the above-described nucleic acid may be of various nature or origin, encompassing proteins of prokaryotic or eukaryotic origin. Among the polypeptides expressed under the control of a *BAP28* regulatory region include bacterial, fungal or viral antigens. Also encompassed are eukaryotic proteins such as intracellular proteins, like "house keeping" proteins, membrane-bound proteins, like receptors, and secreted proteins like endogenous mediators such as cytokines. The desired polypeptide may be the BAP28 protein, especially the protein of the amino acid sequence of SEQ ID No 1, or a fragment or a variant thereof.

The desired nucleic acids encoded by the above-described polynucleotide, usually an RNA molecule, may be complementary to a desired coding polynucleotide, for example to the *BAP28* coding sequence, and thus useful as an antisense polynucleotide.

Such a polynucleotide may be included in a recombinant expression vector in order to express the desired polypeptide or the desired nucleic acid in host cell or in a host organism. Suitable recombinant vectors that contain a polynucleotide such as described hereinbefore are disclosed elsewhere in the specification.

**Polynucleotide Constructs**

The terms "polynucleotide construct" and "recombinant polynucleotide" are used interchangeably herein to refer to linear or circular, purified or isolated polynucleotides that have been artificially designed and which comprise at least two nucleotide sequences that are not found as contiguous nucleotide sequences in their initial natural environment.

**DNA Construct That Enables Directing Temporal And Spatial BAP28 Gene Expression In Recombinant Cell Hosts And In Transgenic Animals.**

In order to study the physiological and phenotypic consequences of a lack of synthesis of the BAP28 protein, both at the cell level and at the multi cellular organism level, the invention also encompasses DNA constructs and recombinant vectors enabling a conditional expression of a specific allele of the *BAP28* genomic sequence or cDNA and also of a copy of this genomic sequence or cDNA harboring substitutions, deletions, or additions of one or more bases as regards to

the *BAP28* nucleotide sequence of SEQ ID Nos 1-3, or a fragment thereof, these base substitutions, deletions or additions being located either in an exon, an intron or a regulatory sequence, but preferably in an exon of the *BAP28* genomic sequence or within the *BAP28* cDNA of SEQ ID No 2 or 3. In a preferred embodiment, the *BAP28* sequence comprises a biallelic marker of the present invention. In a preferred embodiment, the *BAP28* sequence comprises a biallelic marker of the present invention, preferably one of the biallelic markers A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53.

In an additional embodiment, the invention concerns a DNA construct comprising an exon of *PCTA-1* selected from the group consisting of exons A, B, C, and D.

The present invention embodies recombinant vectors comprising any one of the polynucleotides described in the present invention. More particularly, the polynucleotide constructs according to the present invention can comprise any of the polynucleotides described in the "Genomic Sequences Of The Human *BAP28* Gene" section, the " *BAP28* cDNA Sequences" section, the "Coding Regions" section, and the "Oligonucleotide Probes And Primers" section.

A first preferred DNA construct is based on the tetracycline resistance operon *tet* from *E. coli* transposon Tn10 for controlling the *BAP28* gene expression, such as described by Gossen et al.(1992, 1995) and Furth et al.(1994). Such a DNA construct contains seven *tet* operator sequences from Tn10 (*tet*op) that are fused to a minimal promoter, said minimal promoter being operably linked to a polynucleotide of interest that codes either for a sense or an antisense oligonucleotide or for a polypeptide, including a BAP28 polypeptide or a peptide fragment thereof. This DNA construct is functional as a conditional expression system for the nucleotide sequence of interest when the same cell also comprises a nucleotide sequence coding for either the wild type (tTA) or the mutant (rTA) repressor fused to the activating domain of viral protein VP16 of herpes simplex virus, placed under the control of a promoter, such as the HCMVIE1 enhancer/promoter or the MMTV-LTR. Indeed, a preferred DNA construct of the invention comprise both the polynucleotide containing the *tet* operator sequences and the polynucleotide containing a sequence coding for the tTA or the rTA repressor.

In a specific embodiment, the conditional expression DNA construct contains the sequence encoding the mutant tetracycline repressor rTA, the expression of the polynucleotide of interest is silent in the absence of tetracycline and induced in its presence.

**DNA Constructs Allowing Homologous Recombination: Replacement Vectors**

A second preferred DNA construct will comprise, from 5'-end to 3'-end: (a) a first nucleotide sequence that is comprised in the *BAP28* genomic sequence; (b) a nucleotide sequence comprising a positive selection marker, such as the marker for neomycine resistance (*neo*); and (c) a

second nucleotide sequence that is comprised in the *BAP28* genomic sequence, and is located on the genome downstream the first *BAP28* nucleotide sequence (a).

In a preferred embodiment, this DNA construct also comprises a negative selection marker located upstream the nucleotide sequence (a) or downstream the nucleotide sequence (c). Preferably, the negative selection marker consists of the thymidine kinase (*tk*) gene (Thomas et al., 1986), the hygromycine beta gene (Te Riele et al., 1990), the *hprt* gene (Van der Lugt et al., 1991; Reid et al., 1990) or the Diphteria toxin A fragment (*Dt-A*) gene (Nada et al., 1993; Yagi et al.1990). Preferably, the positive selection marker is located within a *BAP28* exon sequence so as to interrupt the sequence encoding a *BAP28* protein. These replacement vectors are described, for example, by Thomas et al.(1986; 1987), Mansour et al.(1988) and Koller et al.(1992).

The first and second nucleotide sequences (a) and (c) may be indifferently located within a *BAP28* regulatory sequence, an intronic sequence, an exon sequence or a sequence containing both regulatory and/or intronic and/or exon sequences. The size of the nucleotide sequences (a) and (c) ranges from 1 to 50 kb, preferably from 1 to 10 kb, more preferably from 2 to 6 kb and most preferably from 2 to 4 kb.

**DNA Constructs Allowing Homologous Recombination: Cre-LoxP System.**

These new DNA constructs make use of the site specific recombination system of the P1 phage. The P1 phage possesses a recombinase called Cre which interacts specifically with a 34 base pairs *lox*P site. The *lox*P site is composed of two palindromic sequences of 13 bp separated by a 8 bp conserved sequence (Hoess et al., 1986). The recombination by the Cre enzyme between two *lox*P sites having an identical orientation leads to the deletion of the DNA fragment.

The Cre-*lox*P system used in combination with a homologous recombination technique has been first described by Gu et al.(1993, 1994). Briefly, a nucleotide sequence of interest to be inserted in a targeted location of the genome harbors at least two *lox*P sites in the same orientation and located at the respective ends of a nucleotide sequence to be excised from the recombinant genome. The excision event requires the presence of the recombinase (Cre) enzyme within the nucleus of the recombinant cell host. The recombinase enzyme may be brought at the desired time either by (a) incubating the recombinant cell hosts in a culture medium containing this enzyme, by injecting the Cre enzyme directly into the desired cell, such as described by Araki et al.(1995), or by lipofection of the enzyme into the cells, such as described by Baubonis et al.(1993); (b) transfecting the cell host with a vector comprising the *Cre* coding sequence operably linked to a promoter functional in the recombinant cell host (in some embodiments, the promoter may be inducible), said vector being introduced in the recombinant cell host, such as described by Gu et al.(1993) and Sauer et al.(1988); (c) introducing in the genome of the cell host a polynucleotide comprising the *Cre* coding sequence operably linked to a promoter functional in the recombinant cell host (in some embodiments, the promoter may be inducible), and said polynucleotide being inserted in the genome

of the cell host either by a random insertion event or an homologous recombination event, such as described by Gu et al.(1994).

In a specific embodiment, the vector containing the sequence to be inserted in the *BAP28* gene by homologous recombination is constructed in such a way that selectable markers are flanked by *lox*P sites of the same orientation, it is possible, by treatment by the Cre enzyme, to eliminate the selectable markers while leaving the *BAP28* sequences of interest that have been inserted by an homologous recombination event. Again, two selectable markers are needed: a positive selection marker to select for the recombination event and a negative selection marker to select for the homologous recombination event. Vectors and methods using the Cre-*lox*P system are described by Zou et al.(1994).

Thus, a third preferred DNA construct of the invention comprises, from 5`-end to 3`-end: (a) a first nucleotide sequence that is comprised in the *BAP28* genomic sequence; (b) a nucleotide sequence comprising a polynucleotide encoding a positive selection marker, said nucleotide sequence comprising additionally two sequences defining a site recognized by a recombinase, such as a *lox*P site, the two sites being placed in the same orientation; and (c) a second nucleotide sequence that is comprised in the *BAP28* genomic sequence, and is located on the genome downstream of the first *BAP28* nucleotide sequence (a).

The sequences defining a site recognized by a recombinase, such as a *lox*P site, are preferably located within the nucleotide sequence (b) at suitable locations bordering the nucleotide sequence for which the conditional excision is sought. In one specific embodiment, two *lox*P sites are located at each side of the positive selection marker sequence, in order to allow its excision at a desired time after the occurrence of the homologous recombination event.

In a preferred embodiment of a method using the third DNA construct described above, the excision of the polynucleotide fragment bordered by the two sites recognized by a recombinase, preferably two loxP sites, is performed at a desired time, due to the presence within the genome of the recombinant host cell of a sequence encoding the Cre enzyme operably linked to a promoter sequence, preferably an inducible promoter, more preferably a tissue-specific promoter sequence and most preferably a promoter sequence which is both inducible and tissue-specific, such as described by Gu et al.(1994).

The presence of the Cre enzyme within the genome of the recombinant cell host may result of the breeding of two transgenic animals, the first transgenic animal bearing the *BAP28*-derived sequence of interest containing the *lox*P sites as described above and the second transgenic animal bearing the *Cre* coding sequence operably linked to a suitable promoter sequence, such as described by Gu et al.(1994).

Spatio-temporal control of the Cre enzyme expression may also be achieved with an adenovirus based vector that contains the Cre gene thus allowing infection of cells, or *in vivo*

infection of organs, for delivery of the Cre enzyme, such as described by Anton and Graham (1995) and Kanegae et al.(1995).

The DNA constructs described above may be used to introduce a desired nucleotide sequence of the invention, preferably a *BAP28* genomic sequence or a *BAP28* cDNA sequence, and most preferably an altered copy of a *BAP28* genomic or cDNA sequence, within a predetermined location of the targeted genome, leading either to the generation of an altered copy of a targeted gene (knock-out homologous recombination) or to the replacement of a copy of the targeted gene by another copy sufficiently homologous to allow an homologous recombination event to occur (knock-in homologous recombination). In a specific embodiment, the DNA constructs described above may be used to introduce a *BAP28* genomic sequence or a *BAP28* cDNA sequence. In some embodiments, said sequence comprises at least one biallelic marker of the present invention, preferably at least one biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53.

**Nuclear Antisense DNA Constructs**

Other compositions containing a vector of the invention comprising an oligonucleotide fragment of the nucleic sequence SEQ ID No 2 or 3, preferably a fragment including the start codon of the *BAP28* gene, as an antisense tool that inhibits the expression of the corresponding *BAP28* gene or the expression of the *PCTA-1* gene. Preferred methods using antisense polynucleotide according to the present invention are the procedures described by Sczakiel et al.(1995) or those described in PCT Application No WO 95/24223.

Preferably, the antisense tools are chosen among the polynucleotides (15-200 bp long) that are complementary to the 5'end or 3' end of the *BAP28* mRNA. In one embodiment, a combination of different antisense polynucleotides complementary to different parts of the desired targeted gene are used.

A preferred antisense according to the invention is a polynucleotide according to the invention, more particularly polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600 or 1000 nucleotides of SEQ ID No 1, or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the following ranges of nucleotide positions of SEQ ID No 1: 57386-27494, 58504-59354, 85947-86108, and 91259-91325.

Preferred antisense polynucleotides according to the present invention are complementary to a sequence of the mRNAs of *BAP28* that contains either the translation initiation codon ATG or a splicing site. Further preferred antisense polynucleotides according to the invention are complementary of the splicing site of the *BAP28* mRNA.

The antisense nucleic acids should have a length and melting temperature sufficient to permit formation of an intracellular duplex having sufficient stability to inhibit the expression of the *BAP28* mRNA in the duplex. Strategies for designing antisense nucleic acids suitable for use in gene therapy are disclosed in Green et al., (1986) and Izant and Weintraub, (1984), the disclosures of which are incorporated herein by reference.

In some strategies, antisense molecules are obtained by reversing the orientation of the *BAP28* coding region with respect to a promoter so as to transcribe the opposite strand from that which is normally transcribed in the cell. The antisense molecules may be transcribed using in vitro transcription systems such as those which employ T7 or SP6 polymerase to generate the transcript. Another approach involves transcription of *BAP28* antisense nucleic acids in vivo by operably linking DNA containing the antisense sequence to a promoter in a suitable expression vector.

Alternatively, suitable antisense strategies are those described by Rossi et al.(1991), in the International Applications Nos. WO 94/23026, WO 95/04141, WO 92/18522 and in the European Patent Application No EP 0 572 287 A2.

Preferably, the antisense polynucleotides of the invention have a 3` polyadenylation signal that has been replaced with a self-cleaving ribozyme sequence, such that RNA polymerase II transcripts are produced without poly(A) at their 3` ends, these antisense polynucleotides being incapable of export from the nucleus, such as described by Liu et al.(1994). In a preferred embodiment, these *BAP28* antisense polynucleotides also comprise, within the ribozyme cassette, a histone stem-loop structure to stabilize cleaved transcripts against 3`-5` exonucleolytic degradation, such as the structure described by Eckner et al.(1991).

An alternative to the antisense technology that is used according to the present invention consists in using ribozymes that will bind to a target sequence via their complementary polynucleotide tail and that will cleave the corresponding RNA by hydrolyzing its target site (namely "hammerhead ribozymes"). Briefly, the simplified cycle of a hammerhead ribozyme consists of (1) sequence specific binding to the target RNA via complementary antisense sequences; (2) site-specific hydrolysis of the cleavable motif of the target strand; and (3) release of cleavage products, which gives rise to another catalytic cycle. Indeed, the use of long-chain antisense polynucleotide (at least 30 bases long) or ribozymes with long antisense arms are advantageous. A preferred delivery system for antisense ribozyme is achieved by covalently linking these antisense ribozymes to lipophilic groups or to use liposomes as a convenient vector. Preferred antisense ribozymes according to the present invention are prepared as described by Sczakiel et al.(1995), the specific preparation procedures being referred to in said article being herein incorporated by reference.

comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the following ranges of nucleotide positions of:

(a) SEQ ID No 1: 1-2500, 2501-5000, 5001-7500, 7501-10000, 10001-12500, 12501-15000, 15001-17500, 17501-20000, 20001-22500, 22501-25000, 25001-27500, 27501-30000, 30001-32500, 32501-35000, 35001-37500, 37501-40000, 40001-42500, 42501-45000, 45001-47500, 47501-50000, 50001-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662;

(b) SEQ ID No 2: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, and 6501 to 6782; and,

(c) SEQ ID No 3: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, 6501 to 7000, 7001 to 7500, 7501 to 7932.

Thus, the invention also relates to nucleic acid probes characterized in that they hybridize specifically, under the stringent hybridization conditions defined above, with a nucleic acid selected from the group consisting of the nucleotide sequences:

a) 1-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662 of SEQ ID No 1 or a variant thereof or a sequence complementary thereto; or

b) 1 to 4995 of SEQ ID No 2 or 3 or a variant thereof or a sequence complementary thereto; and,

c) at least one of nucleotide ranges 1 to 2033, 2160 to 2348, 2676 to 4995 of SEQ ID No 2 or 3, or a variant thereof or a sequence complementary thereto.

Additionally, another preferred embodiment of a probe according to the invention consists of a nucleic acid comprising a biallelic marker selected from the group consisting of A1 to A58 or the complements thereto, for which the respective locations in the sequence listing are provided in Table 2. Preferably, a probe according to the present invention consists of a nucleic acid comprising one of the biallelic markers A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58. More preferably, a probe according to the present invention consists of a nucleic acid comprising one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53.

In one embodiment the invention encompasses isolated, purified, and recombinant polynucleotides comprising, consisting of, or consisting essentially of a contiguous span of 8 to 50 nucleotides of SEQ ID Nos 1, 2, or 3 and the complement thereof, wherein said span includes a *BAP28*-related biallelic marker in said sequence; In some embodiments said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the

biallelic markers in linkage disequilibrium therewith; In some embodiments said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof or the biallelic markers in linkage disequilibrium therewith; In some embodiments said contiguous span is 18 to 35 nucleotides in length and said biallelic marker is within 4 nucleotides of the center of said polynucleotide; In some embodiments, said polynucleotide consists of said contiguous span and said contiguous span is 25 nucleotides in length and said biallelic marker is at the center of said polynucleotide; In some embodiments, the 3' end of said contiguous span is present at the 3' end of said polynucleotide; In some embodiments, the 3' end of said contiguous span is located at the 3' end of said polynucleotide and said biallelic marker is present at the 3' end of said polynucleotide. In a preferred embodiment, said probes comprises, consists of, or consists essentially of a sequence selected from the following sequences: P1 to P58, preferably P1 to P27, P34, P37 to P41, P43 to P49, P52, and P54 to P58, and the complementary sequences thereto.

In another embodiment the invention encompasses isolated, purified and recombinant polynucleotides comprising, consisting of, or consisting essentially of a contiguous span of 8 to 50 nucleotides of SEQ ID Nos 1, 2, or 3 or the complements thereof, wherein the 3' end of said contiguous span is located at the 3' end of said polynucleotide, and wherein the 3' end of said polynucleotide is located within 20 nucleotides upstream of a *BAP28*-related biallelic marker in said sequence; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; optionally, In some embodiments, the 3' end of said polynucleotide is located 1 nucleotide upstream of said *BAP28*-related biallelic marker in said sequence; In some embodiments, said polynucleotide consists essentially of a sequence selected from the following sequences: D1 to D58 and E1 to E58, preferably D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58.

In a further embodiment, the invention encompasses isolated, purified, or recombinant polynucleotides comprising, consisting of, or consisting essentially of a sequence selected from the following sequences: B1 to B38 and C1 to C38, preferably B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, and C34 to C38.

**Oligonucleotide Probes And Primers**

Polynucleotides derived from the *BAP28* gene are useful in order to detect the presence of at least a copy of a nucleotide sequence of SEQ ID Nos 1-3, or a fragment, complement, or variant thereof in a test sample.

Preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, or 200 nucleotides, to the extent that such a length is consistent with the lengths of the particular nucleotide position, of SEQ ID No 1 or the complement thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 20, 30, 40 or 50 nucleotides selected from the group consisting of the following nucleotide positions of SEQ ID No 1: 4997-5076, 5371-5544, 6121-6337, 9877-10018, 11522-11623, 12521-12661, 13453-13664, 13824-13957, 15376-15478, 16855 16965, 17378-17495, 18535-18642, 21446-21541, 21999-22087, 23036-23247, 23546-23667, 24270-24461, 26287-26470, 26611-26747, 28068-28260, 32540-32709, 33112-33270, 34586-34828, 35156-35287, 36660-36763, 36934-37077, 37803-37921, 38017-38138, 40365-40493, 42618-42848, 43452-43578, 44836-44999, 48223-48269, and 49656-49779. Particularly preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 a nucleotide of SEQ ID No 1 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 1: 1-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662.

Particularly preferred embodiments of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of nucleotide positions 1 to 4995 of SEQ ID No 2 or 3. Further embodiments of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 2 or 3: 1 to 2033, 2160 to 2348, and 2676 to 4995.

Additional preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 1-3, or the complements thereof, wherein said contiguous span

In an additional embodiment, the invention encompasses polynucleotides for use in hybridization assay, sequencing assays, and enzyme-based mismatch detection assays for determining the identity of the nucleotide at a *BAP28*-related biallelic marker in SEQ ID No 1, or the complements thereof, as well as polynucleotides for use in amplifying segments of nucleotides comprising a *BAP28*-related biallelic marker in SEQ ID No 1 or the complements thereof; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith.

Furthermore, the present invention also concerns the use of the oligonucleotide probes and primers according to the invention for determining the identity of the nucleotide at a *BAP28*-related biallelic marker. The use of these oligonucleotides in diagnostic is contemplated.

The formation of stable hybrids depends on the melting temperature (Tm) of the DNA. The Tm depends on the length of the primer or probe, the ionic strength of the solution and the G+C content. The higher the G+C content of the primer or probe, the higher is the melting temperature because G:C pairs are held by three H bonds whereas A:T pairs have only two. The GC content in the probes of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

A probe or a primer according to the invention has between 8 and 1000 nucleotides in length, or is specified to be at least 12, 15, 18, 20, 25, 35, 40, 50, 60, 70, 80, 100, 250, 500 or 1000 nucleotides in length. More particularly, the length of these probes and primers can range from 8, 10, 15, 20, or 30 to 100 nucleotides, preferably from 10 to 50, more preferably from 15 to 30 nucleotides. Shorter probes and primers tend to lack specificity for a target nucleic acid sequence and generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. Longer probes and primers are expensive to produce and can sometimes self-hybridize to form hairpin structures. The appropriate length for primers and probes under a particular set of assay conditions may be empirically determined by one of skill in the art. A preferred probe or primer consists of a nucleic acid comprising a polynucleotide selected from the group of the nucleotide sequences of P1 to P58 and the complementary sequences thereto, B1 to B38 and C1 to C38, D1 to D58, E1 to E58, for which the respective locations in the sequence listing are provided in Tables 1, 3, and 4, preferably a nucleic acid comprising a polynucleotide selected from the group of the nucleotide sequences of P1 to P27, P34, P37 to P41, P43 to P49, P52, and P54 to P58, and the complementary sequences thereto, B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to

C15, C22, C24, C25, C27 to 29, C32, C34 to C38, D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58.

The primers and probes can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphodiester method of Narang et al.(1979), the phosphodiester method of Brown et al.(1979), the diethylphosphoramidite method of Beaucage et al.(1981) and the solid support method described in EP 0 707 592. The disclosures of all these documents are incorporated herein by reference.

Detection probes are generally nucleic acid sequences or uncharged nucleic acid analogs such as, for example peptide nucleic acids which are disclosed in International Patent Application WO 92/20702, morpholino analogs which are described in U.S. Patents Numbered 5,185,444; 5,034,506 and 5,142,047. The probe may have to be rendered "non-extendable" in that additional dNTPs cannot be added to the probe. In and of themselves analogs usually are non-extendable and nucleic acid probes can be rendered non-extendable by modifying the 3' end of the probe such that the hydroxyl group is no longer capable of participating in elongation. For example, the 3' end of the probe can be functionalized with the capture or detection label to thereby consume or otherwise block the hydroxyl group. Alternatively, the 3' hydroxyl group simply can be cleaved, replaced or modified. U.S. Patent Application Serial No 07/049,061 filed April 19, 1993 describes modifications, which can be used to render a probe non-extendable.

Any of the polynucleotides of the present invention can be labeled, if desired, by incorporating a label detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include radioactive substances ($^{32}P$, $^{35}S$, $^{3}H$, $^{125}I$), fluorescent dyes (5-bromodesoxyuridin, fluorescein, acetylaminofluorene, digoxigenin) or biotin. Preferably, polynucleotides are labeled at their 3` and 5` ends. Examples of non-radioactive labeling of nucleic acid fragments are described in the French patent No FR-7810975 or by Urdea et al (1988) or Sanchez-Pescador et al (1988). In addition, the probes according to the present invention may have structural characteristics such that they allow the signal amplification, such structural characteristics being, for example, branched DNA probes as those described by Urdea et al. in 1991 or in the European patent No EP 0 225 807 (Chiron).

A label can also be used to capture the primer, so as to facilitate the immobilization of either the primer or a primer extension product, such as amplified DNA, on a solid support. A capture label is attached to the primers or probes and can be a specific binding member which forms a binding pair with the solid's phase reagent's specific binding member (e.g. biotin and streptavidin). Therefore depending upon the type of label carried by a polynucleotide or a probe, it may be employed to capture or to detect the target DNA. Further, it will be understood that the polynucleotides, primers or probes provided herein, may, themselves, serve as the capture label. For example, in the case where a solid phase reagent's binding member is a nucleic acid sequence, it may be selected such that it binds a complementary portion of a primer or probe to thereby

immobilize the primer or probe to the solid phase. In cases where a polynucleotide probe itself serves as the binding member, those skilled in the art will recognize that the probe will contain a sequence or "tail" that is not complementary to the target. In the case where a polynucleotide primer itself serves as the capture label, at least a portion of the primer will be free to hybridize with a nucleic acid on a solid phase. DNA Labeling techniques are well known to the skilled technician.

The probes of the present invention are useful for a number of purposes. They can be notably used in Southern hybridization to genomic DNA. The probes can also be used to detect PCR amplification products. They may also be used to detect mismatches in the *BAP28* gene or mRNA using other techniques.

Any of the polynucleotides, primers and probes of the present invention can be conveniently immobilized on a solid support. Solid supports are known to those skilled in the art and include the walls of wells of a reaction tray, test tubes, polystyrene beads, magnetic beads, nitrocellulose strips, membranes, microparticles such as latex particles, sheep (or other animal) red blood cells, duracytes and others. The solid support is not critical and can be selected by one skilled in the art. Thus, latex particles, microparticles, magnetic or non-magnetic beads, membranes, plastic tubes, walls of microtiter wells, glass or silicon chips, sheep (or other suitable animal's) red blood cells and duracytes are all suitable examples. Suitable methods for immobilizing nucleic acids on solid phases include ionic, hydrophobic, covalent interactions and the like. A solid support, as used herein, refers to any material which is insoluble, or can be made insoluble by a subsequent reaction. The solid support can be chosen for its intrinsic ability to attract and immobilize the capture reagent. Alternatively, the solid phase can retain an additional receptor which has the ability to attract and immobilize the capture reagent. The additional receptor can include a charged substance that is oppositely charged with respect to the capture reagent itself or to a charged substance conjugated to the capture reagent. As yet another alternative, the receptor molecule can be any specific binding member which is immobilized upon (attached to) the solid support and which has the ability to immobilize the capture reagent through a specific binding reaction. The receptor molecule enables the indirect binding of the capture reagent to a solid support material before the performance of the assay or during the performance of the assay. The solid phase thus can be a plastic, derivatized plastic, magnetic or non-magnetic metal, glass or silicon surface of a test tube, microtiter well, sheet, bead, microparticle, chip, sheep (or other suitable animal's) red blood cells, duracytes® and other configurations known to those of ordinary skill in the art. The polynucleotides of the invention can be attached to or immobilized on a solid support individually or in groups of at least 2, 5, 8, 10, 12, 15, 20, or 25 distinct polynucleotides of the invention to a single solid support. In addition, polynucleotides other than those of the invention may be attached to the same solid support as one or more polynucleotides of the invention.

Consequently, the invention also deals with a method for detecting the presence of a nucleic acid comprising a nucleotide sequence selected from a group consisting of SEQ ID Nos 1-4,

9-13, a fragment or a variant thereof and a complementary sequence thereto in a sample, said method comprising the following steps of:

a) bringing into contact a nucleic acid probe or a plurality of nucleic acid probes which can hybridize with a nucleotide sequence included in a nucleic acid selected form the group consisting of the nucleotide sequences of SEQ ID Nos 1-4, 9-13, a fragment or a variant thereof and a complementary sequence thereto and the sample to be assayed; and

b) detecting the hybrid complex formed between the probe and a nucleic acid in the sample.

The invention further concerns a kit for detecting the presence of a nucleic acid comprising a nucleotide sequence selected from a group consisting of SEQ ID Nos 1-4, 9-13, a fragment or a variant thereof and a complementary sequence thereto in a sample, said kit comprising:

a) a nucleic acid probe or a plurality of nucleic acid probes which can hybridize with a nucleotide sequence included in a nucleic acid selected form the group consisting of the nucleotide sequences of SEQ ID Nos 1-4, 9-13, a fragment or a variant thereof and a complementary sequence thereto; and

b) in some embodiments, the kit also comprises reagents necessary for performing the hybridization reaction.

In a first preferred embodiment of this detection method and kit, said nucleic acid probe or the plurality of nucleic acid probes are labeled with a detectable molecule. In a second preferred embodiment of said method and kit, said nucleic acid probe or the plurality of nucleic acid probes has been immobilized on a substrate In a third preferred embodiment, the nucleic acid probe or the plurality of nucleic acid probes comprise either a sequence which is selected from the group consisting of the nucleotide sequences of P1 to P58 and the complementary sequences thereto, B1 to B38, C1 to C38, D1 to D58, E1 to E58 or a biallelic marker selected from the group consisting of A1 to A58 and the complements thereto, preferably a nucleic acid comprising a polynucleotide selected from the group of the nucleotide sequences of P1 to P27, P34, P37 to P41, P43 to P49, P52, and P54 to P58, and the complementary sequences thereto, B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, C34 to C38, D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58, or a biallelic marker selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof.

**Oligonucleotide Arrays**

A substrate comprising a plurality of oligonucleotide primers or probes of the invention may be used either for detecting or amplifying targeted sequences in the *BAP28* gene and may also be used for detecting mutations in the coding or in the non-coding sequences of the *BAP28* gene.

Any polynucleotide provided herein may be attached in overlapping areas or at random locations on the solid support. Alternatively the polynucleotides of the invention may be attached in an ordered array wherein each polynucleotide is attached to a distinct region of the solid support which does not overlap with the attachment site of any other polynucleotide. Preferably, such an ordered array of polynucleotides is designed to be "addressable" where the distinct locations are recorded and can be accessed as part of an assay procedure. Addressable polynucleotide arrays typically comprise a plurality of different oligonucleotide probes that are coupled to a surface of a substrate in different known locations. The knowledge of the precise location of each polynucleotides location makes these "addressable" arrays particularly useful in hybridization assays. Any addressable array technology known in the art can be employed with the polynucleotides of the invention. One particular embodiment of these polynucleotide arrays is known as the Genechips™, and has been generally described in US Patent 5,143,854; PCT publications WO 90/15070 and 92/10092. These arrays may generally be produced using mechanical synthesis methods or light directed synthesis methods which incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis (Fodor et al., 1991). The immobilization of arrays of oligonucleotides on solid supports has been rendered possible by the development of a technology generally identified as "Very Large Scale Immobilized Polymer Synthesis" (VLSIPS™) in which, typically, probes are immobilized in a high density array on a solid surface of a chip. Examples of VLSIPS™ technologies are provided in US Patents 5,143,854; and 5,412,087 and in PCT Publications WO 90/15070, WO 92/10092 and WO 95/11995, which describe methods for forming oligonucleotide arrays through techniques such as light-directed synthesis techniques. In designing strategies aimed at providing arrays of nucleotides immobilized on solid supports, further presentation strategies were developed to order and display the oligonucleotide arrays on the chips in an attempt to maximize hybridization patterns and sequence information. Examples of such presentation strategies are disclosed in PCT Publications WO 94/12305, WO 94/11530, WO 97/29212 and WO 97/31256.

In another embodiment of the oligonucleotide arrays of the invention, an oligonucleotide probe matrix may advantageously be used to detect mutations occurring in the *BAP28* gene and in its regulatory region. For this particular purpose, probes are specifically designed to have a nucleotide sequence allowing their hybridization to the genes that carry known mutations (either by deletion, insertion or substitution of one or several nucleotides). By known mutations, it is meant, mutations on the *BAP28* gene that have been identified according, for example to the technique used by Huang et al.(1996) or Samson et al.(1996).

Another technique that is used to detect mutations in the *BAP28* gene is the use of a high-density DNA array. Each oligonucleotide probe constituting a unit element of the high density DNA array is designed to match a specific subsequence of the *BAP28* genomic DNA or cDNA. Thus, an array consisting of oligonucleotides complementary to subsequences of the target gene sequence is

used to determine the identity of the target sequence with the wild gene sequence, measure its amount, and detect differences between the target sequence and the reference wild gene sequence of the *BAP28* gene. In one such design, termed 4L tiled array, is implemented a set of four probes (A, C, G, T), preferably 15-nucleotide oligomers. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. Consequently, a nucleic acid target of length L is scanned for mutations with a tiled array containing 4L probes, the whole probe set containing all the possible mutations in the known wild reference sequence. The hybridization signals of the 15-mer probe set tiled array are perturbed by a single base change in the target sequence. As a consequence, there is a characteristic loss of signal or a "footprint" for the probes flanking a mutation position. This technique was described by Chee et al. in 1996, which is herein incorporated by reference.

Consequently, the invention concerns an array of nucleic acid molecules comprising at least one polynucleotide described above as probes and primers. Preferably, the invention concerns an array of nucleic acid comprising at least two polynucleotides described above as probes and primers.

A further object of the invention consists of an array of nucleic acid sequences comprising either at least one of the sequences selected from the group consisting of P1 to P58, B1 to B38, C1 to C38, D1 to D58, E1 to E58, the sequences complementary thereto, a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 30, or 40 consecutive nucleotides thereof, or at least one sequence comprising a biallelic marker selected from the group consisting of A1 to A58 and the complements thereto. preferably either at least one of the sequences selected from the group consisting of P1 to P27, P34, P37 to P41, P43 to P49, P52, P54 to P58, B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, C34 to C38, D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58, or at least one sequence comprising a biallelic marker selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof.

The invention also pertains to an array of nucleic acid sequences comprising either at least two of the sequences selected from the group consisting of P1 to P58, B1 to B38, C1 to C38, D1 to D58, E1 to E58, the sequences complementary thereto, a fragment thereof of at least 8 consecutive nucleotides thereof, or at least two sequences comprising a biallelic marker selected from the group consisting of A1 to A58 and the complements thereof. preferably either at least two of the sequences selected from the group consisting of P1 to P27, P34, P37 to P41, P43 to P49, P52, P54 to P58, B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, C34 to C38, D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58 or at least two sequences comprising a biallelic marker selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof.

**Amplification of the *BAP28* gene.**

**1. DNA extraction**

As for the source of the genomic DNA to be subjected to analysis, any test sample can be foreseen without any particular limitation. These test samples include biological samples which can be tested by the methods of the present invention described herein and include human and animal body fluids such as whole blood, serum, plasma, cerebrospinal fluid, urine, lymph fluids, and various external secretions of the respiratory, intestinal and genitourinary tracts, tears, saliva, milk, white blood cells, myelomas and the like; biological fluids such as cell culture supernatants; fixed tissue specimens including tumor and non-tumor tissue and lymph node tissues; bone marrow aspirates and fixed cell specimens. The preferred source of genomic DNA used in the context of the present invention is from peripheral venous blood of each donor.

The techniques of DNA extraction are well-known to the skilled technician. Such techniques are described notably by Mackey et al. (1998).

**2. DNA amplification**

DNA amplification techniques are well-known to those skilled in the art. Amplification techniques that can be used in the context of the present invention include, but are not limited to, the ligase chain reaction (LCR) described in EP-A- 320 308, WO 9320227 and EP-A-439 182, the disclosures of which are incorporated herein by reference, the polymerase chain reaction (PCR, RT-PCR) and techniques such as the nucleic acid sequence based amplification (NASBA) described in Guatelli JC, et al. (1990) and in Compton J. (1991), Q-beta amplification as described in European Patent Application no 4544610, strand displacement amplification as described in Walker et al. (1996) and EP A 684 315 and, target mediated amplification as described in PCT Publication WO 9322461, the disclosure of which is incorporated herein by reference.

LCR and Gap LCR are exponential amplification techniques, both depend on DNA ligase to join adjacent primers annealed to a DNA molecule. In Ligase Chain Reaction (LCR), probe pairs are used which include two primary (first and second) and two secondary (third and fourth) probes, all of which are employed in molar excess to target. The first probe hybridizes to a first segment of the target strand and the second probe hybridizes to a second segment of the target strand, the first and second segments being contiguous so that the primary probes abut one another in 5' phosphate-3'hydroxyl relationship, and so that a ligase can covalently fuse or ligate the two probes into a fused product. In addition, a third (secondary) probe can hybridize to a portion of the first probe and a fourth (secondary) probe can hybridize to a portion of the second probe in a similar abutting fashion. Of course, if the target is initially double stranded, the secondary probes also will hybridize to the target complement in the first instance. Once the ligated strand of primary probes is separated from the target strand, it will hybridize with the third and fourth probes which can be ligated to form a complementary, secondary ligated product. It is important to realize that the ligated products are

functionally equivalent to either the target or its complement. By repeated cycles of hybridization and ligation, amplification of the target sequence is achieved. A method for multiplex LCR has also been described (WO 9320227). Gap LCR (GLCR) is a version of LCR where the probes are not adjacent but are separated by 2 to 3 bases.

For amplification of mRNAs, it is within the scope of the present invention to reverse transcribe mRNA into cDNA followed by polymerase chain reaction (RT-PCR); or, to use a single enzyme for both steps as described in U.S. Patent No 5,322,770 or, to use Asymmetric Gap LCR (RT-AGLCR) as described by Marshall et al. (1994). AGLCR is a modification of GLCR that allows the amplification of RNA.

The PCR technology is the preferred amplification technique used in the present invention. A variety of PCR techniques are familiar to those skilled in the art. For a review of PCR technology, see White (1997) and the publication entitled "PCR Methods and Applications" (1991, Cold Spring Harbor Laboratory Press). In each of these PCR procedures, PCR primers on either side of the nucleic acid sequences to be amplified are added to a suitably prepared nucleic acid sample along with dNTPs and a thermostable polymerase such as Taq polymerase, Pfu polymerase, or Vent polymerase. The nucleic acid in the sample is denatured and the PCR primers are specifically hybridized to complementary nucleic acid sequences in the sample. The hybridized primers are extended. Thereafter, another cycle of denaturation, hybridization, and extension is initiated. The cycles are repeated multiple times to produce an amplified fragment containing the nucleic acid sequence between the primer sites. PCR has further been described in several patents including US Patents 4,683,195, 4,683,202 and 4,965,188. Each of these publications is incorporated by reference.

One of the aspects of the present invention is a method for the amplification of the human *BAP28* gene, particularly of the genomic sequences of SEQ ID No 1 or of the cDNA sequence of SEQ ID No 2, or a fragment or a variant thereof in a test sample, preferably using the PCR technology. The method comprises the steps of contacting a test sample suspected of containing the target *BAP28* encoding sequence or portion thereof with amplification reaction reagents comprising a pair of amplification primers, and eventually in some instances a detection probe that can hybridize with an internal region of amplicon sequences to confirm that the desired amplification reaction has taken place.

Thus, the present invention also relates to a method for the amplification of a human *BAP28* gene sequence, particularly of a portion of the genomic sequences of SEQ ID No 1 or of the cDNA sequence of SEQ ID No 2, 3 or 4, or a variant thereof in a test sample, said method comprising the steps of:

a) contacting a test sample suspected of containing the targeted *BAP28* gene sequence comprised in a nucleotide sequence selected from a group consisting of SEQ ID Nos 1-4, or fragments or variants thereof with amplification reaction reagents comprising a pair of amplification

primers as described above and located on either side of the polynucleotide region to be amplified; and

b) in some embodiments, the method also comprises detecting the amplification products.

The invention also concerns a kit for the amplification of a human *BAP28* gene sequence, particularly of a portion of the genomic sequence of SEQ ID No 1 or of the cDNA sequence of SEQ ID No 2, 3 or 4, or a variant thereof in a test sample, wherein said kit comprises:

a) a pair of oligonucleotide primers located on either side of the *BAP28* region to be amplified; and

b) in some embodiments, the kit also comprises the reagents necessary for performing the amplification reaction.

In a first preferred embodiment of the above amplification method or kit, the amplification product is detected by hybridization with a labeled probe having a sequence which is complementary to the amplified region. In another embodiment of the above amplification method and kit, primers comprise a sequence which is selected from the group consisting of the nucleotide sequences of B1 to B38, C1 to C38, D1 to D58, and E1 to E58, preferably B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, C34 to C38, D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58

The primers are more particularly characterized in that they have sufficient complementarity with any sequence of a strand of the genomic sequence close to the region to be amplified, for example with a non-coding sequence adjacent to exons to amplify.

**BAP28 Proteins and Polypeptide Fragments:**

The BAP28 protein has 2144 amino acids in length. This protein is highly conserved in various species such as *Drosophila melanogaster*, *Arabidopsis thaliana*, *Schizosaccahromyces pombe*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Tetraodon nigroviridis*. The protein alignment between the human BAP28 and the proteins from *Drosophila melanogaster*, *Arabidopsis thaliana*, *Schizosaccahromyces pombe*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* is disclosed in the Figure 3. The protein alignment between the human BAP28 and the protein from *Tetraodon nigroviridis* is disclosed in the Figure 4. The BAP28 protein is also well conserved among the mammalian. Indeed, several ESTs with a good homolgy with the human BAP28 have been identified. Some examples of ESTs are the following (Genbank Accession Number/species) : AW423202/zebrafish ; AW481398/Bos taurus ; AW325866/Bos taurus ; AW353291/Bos taurus ; AW315340/Bos taurus ; AA681616/mouse ; AV120680/Mus musculus ; and, D77458/ mouse.

Analysis of the BAP28 protein sequence provided several potential phosphorylation sites and N-glycolsylation sites in BAP28. More particularly, protein kinase C phosphorylation sites have been identified in positions 199-201, 269-271, 387-389, 415-417, 508-510, 650-652, 717-719, 778-780, 792-794, 884-886, 903-905, 999-1001, 1091-1093, 1349-1351, 1506-1508, 1573-1575, 1614-1616, 1632-1634, 1673-1675, 1743-1745, 1808-1810, 1829-1831, 1911-1913, and 2077-2079 of

SEQ ID No4: casein kinase II phosphorylation sites have been identified in positions 22-25, 50-53, 253-256, 363-366, 408-411, 409-412, 508-511, 539-542, 590-593, 689-692, 717-720, 745-748, 961-964, 979-982, 1091-1094, 1105-1108, 1195-1198, 1492-1495, 1723-1726, 1882-1885, 1972-1975, and 1981-1984 of SEQ ID No4. Otherwise, several potential N-glycosylation sites have been identified in positions 93-96, 154-157, 776-779, 882-885, 1347-1350, 1488-1491, 1630-1633, 1746-1749, and 1970-1973 of SEQ ID No 5. A conserved HEAT_REPEAT motif has been identified in positions 2106-2139 of SEQ ID No 5. The HEAT_REPEAT motif are generally involved in protein-protein interaction. The PCT application WO98/12327 showed that BAP28 should be involved in interaction with BRCA1.

The term "BAP28 polypeptides" is used herein to embrace all of the proteins and polypeptides of the present invention. Also forming part of the invention are polypeptides encoded by the polynucleotides of the invention, as well as fusion polypeptides comprising such polypeptides. The invention embodies BAP28 proteins from humans, including isolated or purified BAP28 proteins consisting, consisting essentially, or comprising the sequence of SEQ ID No 5 or fragments thereof. The present invention also embodies isolated, purified, and recombinant polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes at least 1, 2, 3, 5 or 10 of the amino acid positions 1 to 1629 of the SEQ ID No 5. The present invention also embodies isolated, purified, and recombinant polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span include an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5. In other preferred embodiments the BAP28 protein contains an alanine residue at amino acid position 2050 in SEQ ID No 5.

Four biallelic markers of the present invention, namely A16, A19, A21 and A25, provide an amino acid sequence change. Indeed, the biallelic marker A16 encodes a Ser or Asn residue at the position 1694 of the BAP28 protein; the biallelic marker A19 encodes a Ala or Val residue at the position 1854 of the BAP28 protein; the biallelic marker A21 encodes a Asp or Asn at the position 1967 of the BAP28 protein; and the biallelic marker A25 encodes a Gly or Glu at the position 2017 of the BAP28 protein. The invention encompasses the BAP28 proteins comprising all the combinations of the above-described residues at the positions 1694, 1854, 1967, and 2017.

The variant protein and fragments thereof which contain an asparagine at the amino acid position 1694 of SEQ ID No 5 are collectively referred to herein as "1694-Asn variants". The variant protein and fragments thereof which contain a valine at the amino acid position 1854 of SEQ ID No

5 are collectively referred to herein as "1854-Val variants". The variant protein and fragments thereof which contain an asparagine at the amino acid position 1967 of SEQ ID No 5 are collectively referred to herein as "1967-Asn variants". The variant protein and fragments thereof which contain a glutamic acid at the amino acid position 2017 of SEQ ID No 5 are collectively referred to herein as "2017-Glu variants". The variant protein and fragments thereof which contain an alanine at the amino acid position 2050 of SEQ ID No 5 are collectively referred to herein as "2050-Ala variants". In other preferred embodiments of the polypeptides of the present invention, the contiguous stretch of amino acids comprises the site of a mutation or functional mutation, including a deletion, addition, swap or truncation of the amino acids in the BAP28 protein sequence.

The invention also encompasses a purified, isolated, or recombinant polypeptide comprising an amino acid sequence having at least 70, 75, 80, 85, 90, 95, 98 or 99% amino acid identity with the amino acid sequence of SEQ ID No 5 or a fragment thereof.

The invention concerns the polypeptide which are encoded by a nucleic acid comprising a sequence selected from the group consisting of the sequence SEQ ID Nos 1-3 or fragments thereof.

BAP28 proteins are preferably isolated from human or mammalian tissue samples or expressed from human or mammalian genes. The BAP28 polypeptides of the invention can be made using routine expression methods known in the art. The polynucleotide encoding the desired polypeptide is ligated into an expression vector suitable for any convenient host. Both eukaryotic and prokaryotic host systems may be used in forming recombinant polypeptides, and a summary of some of the more common systems. The polypeptide is then isolated from lysed cells or from the culture medium and purified to the extent needed for its intended use. Purification is by any technique known in the art, for example, differential extraction, salt fractionation, chromatography, centrifugation, and the like. See, for example, Methods in Enzymology for a variety of methods for purifying proteins.

In addition, shorter protein fragments is produced by chemical synthesis. Alternatively the proteins of the invention is extracted from cells or tissues of humans or non-human animals. Methods for purifying proteins are known in the art, and include the use of detergents or chaotropic agents to disrupt particles followed by differential extraction and separation of the polypeptides by ion exchange chromatography, affinity chromatography, sedimentation according to density, and gel electrophoresis.

Any *BAP28* cDNA, including SEQ ID Nos 2 and 3, or fragments thereof is used to express BAP28 proteins and polypeptides. The nucleic acid encoding the BAP28 protein or fragments thereof to be expressed is operably linked to a promoter in an expression vector using conventional cloning technology. The *BAP28* insert in the expression vector may comprise the full coding sequence for the BAP28 protein or a portion thereof. For example, the *BAP28* derived insert may encode a polypeptide comprising at least 10 consecutive amino acids of the BAP28 protein of SEQ ID No 5, wherein said

contiguous span includes at least 1, 2, 3, 5 or 10 of the amino acid positions 1 to 1629 of the SEQ ID No 5, or wherein polypeptide is a 2050-Ala variant BAP28 polypeptide.

The expression vector is any of the mammalian, yeast, insect or bacterial expression systems known in the art. Commercially available vectors and expression systems are available from a variety of suppliers including Genetics Institute (Cambridge, MA), Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence is optimized for the particular expression organism in which the expression vector is introduced, as explained by Hatfield, et al., U.S. Patent No 5,082,767.

In one embodiment, the entire coding sequence of the *BAP28* cDNA through the poly A signal of the cDNA are operably linked to a promoter in the expression vector. Alternatively, if the nucleic acid encoding a portion of the BAP28 protein lacks a methionine to serve as the initiation site, an initiating methionine can be introduced next to the first codon of the nucleic acid using conventional techniques. Similarly, if the insert from the *BAP28* cDNA lacks a poly A signal, this sequence can be added to the construct by, for example, splicing out the Poly A signal from pSG5 (Stratagene) using BglI and SalI restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene). pXT1 contains the LTRs and a portion of the gag gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex Thymidine Kinase promoter and the selectable neomycin gene. The nucleic acid encoding the BAP28 protein or a portion thereof is obtained by PCR from a bacterial vector containing the *BAP28* cDNA of SEQ ID No 2 or 3 using oligonucleotide primers complementary to the *BAP28* cDNA or portion thereof and containing restriction endonuclease sequences for Pst I incorporated into the 5'primer and BglII at the 5' end of the corresponding cDNA 3' primer, taking care to ensure that the sequence encoding the BAP28 protein or a portion thereof is positioned properly with respect to the poly A signal. The purified fragment obtained from the resulting PCR reaction is digested with PstI, blunt ended with an exonuclease, digested with Bgl II, purified and ligated to pXT1, now containing a poly A signal and digested with BglII.

The ligated product is transfected into mouse NIH 3T3 cells using Lipofectin (Life Technologies, Inc., Grand Island, New York) under conditions outlined in the product specification. Positive transfectants are selected after growing the transfected cells in 600ug/ml G418 (Sigma, St. Louis, Missouri).

Alternatively, the nucleic acids encoding the BAP28 protein or a portion thereof is cloned into pED6dpc2 (Genetics Institute, Cambridge, MA). The resulting pED6dpc2 constructs is transfected into a suitable host cell, such as COS 1 cells. Methotrexate resistant cells are selected and expanded.

The above procedures may also be used to express a mutant BAP28 protein responsible for a detectable phenotype or a portion thereof.

The expressed proteins are purified using conventional purification techniques such as ammonium sulfate precipitation or chromatographic separation based on size or charge. The protein encoded by the nucleic acid insert may also be purified using standard immunochromatography techniques. In such procedures, a solution containing the expressed BAP28 protein or portion thereof, such as a cell extract, is applied to a column having antibodies against the BAP28 protein or portion thereof is attached to the chromatography matrix. The expressed protein is allowed to bind the immunochromatography column. Thereafter, the column is washed to remove non-specifically bound proteins. The specifically bound expressed protein is then released from the column and recovered using standard techniques.

To confirm expression of the BAP28 protein or a portion thereof, the proteins expressed from host cells containing an expression vector containing an insert encoding the BAP28 protein or a portion thereof can be compared to the proteins expressed in host cells containing the expression vector without an insert. The presence of a band in samples from cells containing the expression vector with an insert which is absent in samples from cells containing the expression vector without an insert indicates that the BAP28 protein or a portion thereof is being expressed. Generally, the band will have the mobility expected for the BAP28 protein or portion thereof. However, the band may have a mobility different than that expected as a result of modifications such as glycosylation, ubiquitination, or enzymatic cleavage.

Antibodies capable of specifically recognizing the expressed BAP28 protein or a portion thereof are described below.

If antibody production is not possible, the nucleic acids encoding the BAP28 protein or a portion thereof is incorporated into expression vectors designed for use in purification schemes employing chimeric polypeptides. In such strategies the nucleic acid encoding the BAP28 protein or a portion thereof is inserted in frame with the gene encoding the other half of the chimera. The other half of the chimera is β-globin or a nickel binding polypeptide encoding sequence. A chromatography matrix having antibody to β-globin or nickel attached thereto is then used to purify the chimeric protein. Protease cleavage sites is engineered between the β-globin gene or the nickel binding polypeptide and the BAP28 protein or portion thereof. Thus, the two polypeptides of the chimera is separated from one another by protease digestion.

One useful expression vector for generating β-globin chimerics is pSG5 (Stratagene), which encodes rabbit β-globin. Intron II of the rabbit β-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal incorporated into the construct increases the level of expression. These techniques are well known to those skilled in the art of molecular biology. Standard methods are published in methods texts such as Davis et al., (1986) and many of the methods are available from Stratagene, Life Technologies, Inc., or Promega. Polypeptide may additionally be produced from the construct using in vitro translation systems such as the In vitro Express™ Translation Kit (Stratagene).

**Antibodies That Bind BAP28 Polypeptides of the Invention**

Any BAP28 polypeptide or whole protein may be used to generate antibodies capable of specifically binding to expressed BAP28 protein or fragments thereof as described. The antibody compositions of the invention are capable of specifically binding or specifically bind to the BAP28 protein. For an antibody composition to specifically bind to the BAP28 protein it must demonstrate at least a 5%, 10%, 15%, 20%, 25%, 50%, or 100% greater binding affinity for full length BAP28 protein than for any full length protein in an ELISA, RIA, or other antibody-based binding assay. For an antibody composition to specifically bind to the 1694-Asn, 1854-Val, 1967-Asn, 2017-Glu, or 2050-Ala variant BAP28 protein, it must demonstrate at least a 5%, 10%, 15%, 20%, 25%, 50%, or 100% greater binding affinity for full length 1694-Asn, 1854-Val, 1967-Asn, 2017-Glu, or 2050-Ala variant BAP28 protein than for respectively a 1694-Ser, 1854-Ala, 1967-Asp, 2017-Gly or 2050-Val full length protein in an ELISA, RIA, or other antibody-based binding assay. The present invention also contemplates the antibodies which are specific of a protein BAP28 comprising one combination of the above-described residues at the positions 1694, 1854, 1967, and 2017.

In a preferred embodiment of the invention antibody compositions are capable of selectively binding, or selectively bind to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said epitope comprises at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 and 2050 of SEQ ID No 5, wherein said antibody composition is optionally either polyclonal or monoclonal. In a other preferred embodiment, antibody compositions are capable of selectively binding, or selectively bind to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said epitope comprises an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, wherein said antibody composition is optionally either polyclonal or monoclonal.

The present invention also contemplates the use of polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 50, or 100 amino acids of a BAP28 polypeptide in the manufacture of antibodies, wherein said contiguous span comprises at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of SEQ ID No 5. The present invention further contemplates the use of polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 50, or 100 amino acids of a BAP28 polypeptide in the manufacture of antibodies, wherein said contiguous span comprises an amino acid

selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5. In a preferred embodiment such polypeptides are useful in the manufacture of antibodies to detect the presence and absence of the BAP28 protein.

Non-human animals or mammals, whether wild-type or transgenic, which express a different species of BAP28 than the one to which antibody binding is desired, and animals which do not express BAP28 (i.e. a BAP28 knock out animal as described in herein) are particularly useful for preparing antibodies. BAP28 knock out animals will recognize all or most of the exposed regions of BAP28 as foreign antigens, and therefore produce antibodies with a wider array of BAP28 epitopes. Moreover, smaller polypeptides with only 10 to 30 amino acids may be useful in obtaining specific binding to the BAP28 protein. In addition, the humoral immune system of animals which produce a species of BAP28 that resembles the antigenic sequence will preferentially recognize the differences between the animal's native BAP28 species and the antigen sequence, and produce antibodies to these unique sites in the antigen sequence. Such a technique will be particularly useful in obtaining antibodies that specifically bind to the BAP28 protein.

Antibody preparations prepared according to either protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample. The antibodies may also be used in therapeutic compositions for killing cells expressing the protein or reducing the levels of the protein in the body.

The antibodies of the invention may be labeled, either by a radioactive, a fluorescent or an enzymatic label.

Consequently, the invention is also directed to a method for detecting specifically the presence of a human BAP28 polypeptide according to the invention in a biological sample, said method comprising the following steps:

a) bringing into contact the biological sample with a polyclonal or monoclonal antibody directed against the BAP28 polypeptide of the amino acid sequence of SEQ ID No 5, or to a peptide fragment or variant thereof;

b) detecting the antigen-antibody complex formed.

The invention also concerns a diagnostic kit for detecting *in vitro* the presence of a human BAP28 polypeptide according to the present invention in a biological sample, wherein said kit comprises :

a) a polyclonal or monoclonal antibody directed against the BAP28 polypeptide of the amino acid sequence of SEQ ID No 5, or to a peptide fragment or variant thereof. In some embodiments, the antibody may be labeled;

b) a reagent allowing the detection of the antigen-antibody complexes formed, said reagent optionally being labelled, or being able to be recognized itself by a labeled reagent, more particularly in the case when the above-mentioned monoclonal or polyclonal antibody is not labeled by itself.

***BAP28* -related Biallelic Markers**

**Advantages Of The Biallelic Markers Of The Present Invention**

The *BAP28*-related biallelic markers of the present invention offer a number of important advantages over other genetic markers such as RFLP (Restriction fragment length polymorphism) and VNTR (Variable Number of Tandem Repeats) markers.

The first generation of markers, were RFLPs, which are variations that modify the length of a restriction fragment. But methods used to identify and to type RFLPs are relatively wasteful of materials, effort, and time. The second generation of genetic markers were VNTRs, which can be categorized as either minisatellites or microsatellites. Minisatellites are tandemly repeated DNA sequences present in units of 5-50 repeats which are distributed along regions of the human chromosomes ranging from 0.1 to 20 kilobases in length. Since they present many possible alleles, their informative content is very high. Minisatellites are scored by performing Southern blots to identify the number of tandem repeats present in a nucleic acid sample from the individual being tested. However, there are only $10^4$ potential VNTRs that can be typed by Southern blotting. Moreover, both RFLP and VNTR markers are costly and time-consuming to develop and assay in large numbers.

Single nucleotide polymorphism or biallelic markers can be used in the same manner as RFLPs and VNTRs but offer several advantages. SNP are densely spaced in the human genome and represent the most frequent type of variation. An estimated number of more than $10^7$ sites are scattered along the $3x10^9$ base pairs of the human genome. Therefore, SNP occur at a greater frequency and with greater uniformity than RFLP or VNTR markers which means that there is a greater probability that such a marker will be found in close proximity to a genetic locus of interest. SNP are less variable than VNTR markers but are mutationally more stable.

Also, the different forms of a characterized single nucleotide polymorphism, such as the biallelic markers of the present invention, are often easier to distinguish and can therefore be typed easily on a routine basis. Biallelic markers have single nucleotide based alleles and they have only two common alleles, which allows highly parallel detection and automated scoring. The biallelic markers of the present invention offer the possibility of rapid, high throughput genotyping of a large number of individuals.

Biallelic markers are densely spaced in the genome, sufficiently informative and can be assayed in large numbers. The combined effects of these advantages make biallelic markers extremely valuable in genetic studies. Biallelic markers can be used in linkage studies in families, in allele sharing methods, in linkage disequilibrium studies in populations, in association studies of

case-control populations or of trait positive and trait negative populations. An important aspect of the present invention is that biallelic markers allow association studies to be performed to identify genes involved in complex traits. Association studies examine the frequency of marker alleles in unrelated case- and control-populations and are generally employed in the detection of polygenic or sporadic traits. Association studies may be conducted within the general population and are not limited to studies performed on related individuals in affected families (linkage studies). Biallelic markers in different genes can be screened in parallel for direct association with disease or response to a treatment. This multiple gene approach is a powerful tool for a variety of human genetic studies as it provides the necessary statistical power to examine the synergistic effect of multiple genetic factors on a particular phenotype, drug response, sporadic trait, or disease state with a complex genetic etiology.

Although most valuable in association studies, the biallelic markers of the present invention can have a wide range of uses, and may for example also be used in forensic identification of individual humans, such as for identification of descendants, determination of paternity, criminal identification, and the like. For example, a DNA sample is obtained from a person or from a cellular sample (e.g. , crime scene evidence such as blood, saliva, semen, and the like) and the identity of the allele present at any one or preferably multiple biallelic markers is determined according to any of the detection methods described herein. On the basis of the allele(s) present at the specified positions, the individual from which the sample originated will be identified with respect to his/her genotype. The biallelic markers of the invention may be used alone or in conjunction with other genetic markers, including RFLP and VNTR to conclusively identify an individual or to rule out the individual as a possible perpetrator.

### *BAP28*-Related Biallelic Markers And Polynucleotides Related Thereto

The invention also concerns *BAP28*-related biallelic markers. A portion of the biallelic markers of the present invention designated A1 to A58 are disclosed in Table 2, including their location on the *BAP28* gene. These biallelic markers are also each listed as a single base polymorphism in the features of SEQ ID No 1.

The invention also relates to a purified and/or isolated nucleotide sequence comprising a polymorphic base of a *BAP28*-related biallelic marker, preferably of a biallelic marker selected from the group consisting of A1 to A58, more preferably one of the biallelic markers A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, still more preferably one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof. The sequence has between 8 and 1000 nucleotides in length, and preferably comprises at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 60, 70, 80, 100, 250, 500 or 1000 contiguous nucleotides of a nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 2 or 3, or a variant thereof or a complementary sequence thereto. These nucleotide sequences comprise the polymorphic base of either allele 1 or

allele 2 of the respective biallelic marker. In some embodiments, said biallelic marker may be within 6, 5, 4, 3, 2, or 1 nucleotides of the center of said polynucleotide or at the center of said polynucleotide. In some embodiments, the 3' end of said contiguous span may be present at the 3' end of said polynucleotide. In some embodiments, a *BAP28*-related biallelic marker biallelic marker may be present at the 3' end of said polynucleotide. In some embodiments, the 3' end of said polynucleotide may be located within or at least 2, 4, 6, 8, 10, 12, 15, 18, 20, 25, 50, 100, 250, 500, or 1000 nucleotides upstream of a *BAP28*-related biallelic marker in said sequence. In some embodiments, the 3' end of said polynucleotide may be located 1 nucleotide upstream of a *BAP28*-related biallelic marker in said sequence. In some embodiments, said polynucleotide may further comprise a label. In some embodiments, said polynucleotide can be attached to solid support. In a further embodiment, the polynucleotides defined above can be used alone or in any combination.

The invention further concerns a nucleic acid encoding the BAP28 protein, wherein said nucleic acid comprises a polymorphic base of a biallelic marker selected from the group consisting of A1 to A58 and the complements thereof, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58.

The invention also encompasses the use of any polynucleotide for, or any polynucleotide for use in, determining the identity of one or more nucleotides at a *BAP28*-related biallelic marker. In addition, the polynucleotides of the invention for use in determining the identity of one or more nucleotides at a *BAP28*-related biallelic marker encompass polynucleotides with any further limitation described in this disclosure, or those following, specified alone or in any combination. In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said polynucleotide may comprise a sequence disclosed in the present specification; In some embodiments, said polynucleotide may comprise, consist of, or consist essentially of any polynucleotide described in the present specification; In some embodiments, said determining may be performed in a hybridization assay, sequencing assay, microsequencing assay, or an enzyme-based mismatch detection assay; In some embodiments, said polynucleotide may be attached to a solid support, array, or addressable array; In some embodiments, said polynucleotide may be labeled. A preferred polynucleotide may be used in a hybridization assay for determining the identity of the nucleotide at a *BAP28*-related biallelic marker. Another preferred polynucleotide may be used in a sequencing or microsequencing assay for determining the identity of the nucleotide at a *BAP28*-related biallelic marker. A third preferred polynucleotide may be used in an enzyme-

based mismatch detection assay for determining the identity of the nucleotide at a *BAP28*-related biallelic marker. A fourth preferred polynucleotide may be used in amplifying a segment of polynucleotides comprising a *BAP28*-related biallelic marker. In some embodiments, any of the polynucleotides described above may be attached to a solid support, array, or addressable array; In some embodiments, said polynucleotide may be labeled.

Additionally, the invention encompasses the use of any polynucleotide for, or any polynucleotide for use in, amplifying a segment of nucleotides comprising a *BAP28*-related biallelic marker. In addition, the polynucleotides of the invention for use in amplifying a segment of nucleotides comprising a *BAP28*-related biallelic marker encompass polynucleotides with any further limitation described in this disclosure, or those following, specified alone or in any combination: In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, wherein said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said polynucleotide may comprise a sequence disclosed in the present specification; In some embodiments, said polynucleotide may comprise, consist of, or consist essentially of any polynucleotide described in the present specification; In some embodiments, said amplifying may be performed by a PCR or LCR. In some embodiments, said polynucleotide may be attached to a solid support, array, or addressable array. In some embodiments, said polynucleotide may be labeled.

The primers for amplification or sequencing reaction of a polynucleotide comprising a biallelic marker of the invention may be designed from the disclosed sequences for any method known in the art. A preferred set of primers are fashioned such that the 3' end of the contiguous span of identity with a sequence selected from the group consisting of SEQ ID Nos 1, 2 or 3, or a sequence complementary thereto or a variant thereof is present at the 3' end of the primer. Such a configuration allows the 3' end of the primer to hybridize to a selected nucleic acid sequence and dramatically increases the efficiency of the primer for amplification or sequencing reactions. Allele specific primers may be designed such that a polymorphic base of a biallelic marker is at the 3' end of the contiguous span and the contiguous span is present at the 3' end of the primer. Such allele specific primers tend to selectively prime an amplification or sequencing reaction so long as they are used with a nucleic acid sample that contains one of the two alleles present at a biallelic marker. The 3' end of the primer of the invention may be located within or at least 2, 4, 6, 8, 10, 12, 15, 18, 20, 25, 50, 100, 250, 500, or 1000 nucleotides upstream of a *BAP28*-related biallelic marker in said sequence or at any other location which is appropriate for their intended use in sequencing.

amplification or the location of novel sequences or markers. Thus, another set of preferred amplification primers comprise an isolated polynucleotide consisting essentially of a contiguous span of 8 to 50 nucleotides in a sequence selected from the group consisting of SEQ ID Nos 1, 2 or 3 or a sequence complementary thereto or a variant thereof, wherein the 3` end of said contiguous span is located at the 3`end of said polynucleotide, and wherein the 3`end of said polynucleotide is located upstream of a *BAP28*-related biallelic marker in said sequence. Preferably, those amplification primers comprise a sequence selected from the group consisting of the sequences B1 to B38 and C1 to C38, preferably B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, and C34 to C38. Primers with their 3' ends located 1 nucleotide upstream of a biallelic marker of *BAP28* have a special utility as microsequencing assays. Preferred microsequencing primers are described in Table 4. In some embodiments, microsequencing primers are selected from the group consisting of the nucleotide sequences D1 to D58 and E1 to E58, preferably D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58.

The probes of the present invention may be designed from the disclosed sequences for any method known in the art, particularly methods which allow for testing if a marker disclosed herein is present. A preferred set of probes may be designed for use in the hybridization assays of the invention in any manner known in the art such that they selectively bind to one allele of a biallelic marker, but not the other under any particular set of assay conditions. Preferred hybridization probes comprise the polymorphic base of either allele 1 or allele 2 of the considered biallelic marker. In some embodiments, said biallelic marker may be within 6, 5, 4, 3, 2, or 1 nucleotides of the center of the hybridization probe or at the center of said probe. In a preferred embodiment, the probes are selected in the group consisting of the sequences P1 to P58 and the complementary sequence thereto (Table 3), preferably P1 to P27, P34, P37 to P41, P43 to P49, P52, and P54 to P58.

It should be noted that the polynucleotides of the present invention are not limited to having the exact flanking sequences surrounding the polymorphic bases which are enumerated in Sequence Listing. Rather, it will be appreciated that the flanking sequences surrounding the biallelic markers may be lengthened or shortened to any extent compatible with their intended use and the present invention specifically contemplates such sequences. The flanking regions outside of the contiguous span need not be homologous to native flanking sequences which actually occur in human subjects. The addition of any nucleotide sequence which is compatible with the nucleotides intended use is specifically contemplated.

Primers and probes may be labeled or immobilized on a solid support as described in "Oligonucleotide probes and primers". The polynucleotides of the invention which are attached to a solid support encompass polynucleotides with any further limitation described in this disclosure, or those following, specified alone or in any combination: In some embodiments, said polynucleotides may be specified as attached individually or in groups of at least 2, 5, 8, 10, 12, 15, 20, or 25 distinct

polynucleotides of the invention to a single solid support. In some embodiments, polynucleotides other than those of the invention may attached to the same solid support as polynucleotides of the invention. In some embodiments, when multiple polynucleotides are attached to a solid support they may be attached at random locations, or in an ordered array. In some embodiments, said ordered array may be addressable.

The present invention also encompasses diagnostic kits comprising one or more polynucleotides of the invention with a portion or all of the necessary reagents and instructions for genotyping a test subject by determining the identity of a nucleotide at a *BAP28*-related biallelic marker. The polynucleotides of a kit may optionally be attached to a solid support, or be part of an array or addressable array of polynucleotides. The kit may provide for the determination of the identity of the nucleotide at a marker position by any method known in the art including, but not limited to, a sequencing assay method, a microsequencing assay method, a hybridization assay method, or an enzyme-based mismatch detection assay method.

**Methods For *De Novo* Identification Of Biallelic Markers**

Any of a variety of methods can be used to screen a genomic fragment for single nucleotide polymorphisms such as differential hybridization with oligonucleotide probes, detection of changes in the mobility measured by gel electrophoresis or direct sequencing of the amplified nucleic acid. A preferred method for identifying biallelic markers involves comparative sequencing of genomic DNA fragments from an appropriate number of unrelated individuals.

In a first embodiment, DNA samples from unrelated individuals are pooled together, following which the genomic DNA of interest is amplified and sequenced. The nucleotide sequences thus obtained are then analyzed to identify significant polymorphisms. One of the major advantages of this method resides in the fact that the pooling of the DNA samples substantially reduces the number of DNA amplification reactions and sequencing reactions, which must be carried out. Moreover, this method is sufficiently sensitive so that a biallelic marker obtained thereby usually demonstrates a sufficient frequency of its less common allele to be useful in conducting association studies.

In a second embodiment, the DNA samples are not pooled and are therefore amplified and sequenced individually. This method is usually preferred when biallelic markers need to be identified in order to perform association studies within candidate genes. Preferably, highly relevant gene regions such as promoter regions or exon regions may be screened for biallelic markers. A biallelic marker obtained using this method may show a lower degree of informativeness for conducting association studies, e.g. if the frequency of its less frequent allele may be less than about 10%. Such a biallelic marker will, however, be sufficiently informative to conduct association studies and it will further be appreciated that including less informative biallelic markers in the genetic analysis studies of the present invention, may allow in some cases the direct identification of causal mutations, which may, depending on their penetrance, be rare mutations.

The following is a description of the various parameters of a preferred method used by the inventors for the identification of the biallelic markers of the present invention.

**Genomic DNA Samples**

The genomic DNA samples from which the biallelic markers of the present invention are generated are preferably obtained from unrelated individuals corresponding to a heterogeneous population of known ethnic background. The number of individuals from whom DNA samples are obtained can vary substantially, preferably from about 10 to about 1000, preferably from about 50 to about 200 individuals. It is usually preferred to collect DNA samples from at least about 100 individuals in order to have sufficient polymorphic diversity in a given population to identify as many markers as possible and to generate statistically significant results.

As for the source of the genomic DNA to be subjected to analysis, any test sample can be foreseen without any particular limitation. These test samples include biological samples, which can be tested by the methods of the present invention described herein, and include human and animal body fluids such as whole blood, serum, plasma, cerebrospinal fluid, urine, lymph fluids, and various external secretions of the respiratory, intestinal and genitourinary tracts, tears, saliva, milk, white blood cells, myelomas and the like; biological fluids such as cell culture supernatants; fixed tissue specimens including tumor and non-tumor tissue and lymph node tissues; bone marrow aspirates and fixed cell specimens. The preferred source of genomic DNA used in the present invention is from peripheral venous blood of each donor. Techniques to prepare genomic DNA from biological samples are well known to the skilled technician. Details of a preferred embodiment are provided in Example 1. The person skilled in the art can choose to amplify pooled or unpooled DNA samples.

**DNA Amplification**

The identification of biallelic markers in a sample of genomic DNA may be facilitated through the use of DNA amplification methods. DNA samples can be pooled or unpooled for the amplification step. DNA amplification techniques are well known to those skilled in the art. Various methods to amplify DNA fragments carrying biallelic markers are further described hereinbefore in "Amplification of the *BAP28* gene". The PCR technology is the preferred amplification technique used to identify new biallelic markers. A typical example of a PCR reaction suitable for the purposes of the present invention is provided in Example 2.

In a first embodiment of the present invention, biallelic markers are identified using genomic sequence information generated by the inventors. Sequenced genomic DNA fragments are used to design primers for the amplification of 500 bp fragments. These 500 bp fragments are amplified from genomic DNA and are scanned for biallelic markers. Primers may be designed using the OSP software (Hillier L. and Green P., 1991). All primers may contain, upstream of the specific

target bases, a common oligonucleotide tail that serves as a sequencing primer. Those skilled in the art are familiar with primer extensions, which can be used for these purposes.

Preferred primers, useful for the amplification of genomic sequences encoding the candidate genes, focus on promoters, exons and splice sites of the genes. A biallelic marker presents a higher probability to be an eventual causal mutation if it is located in these functional regions of the gene. Preferred amplification primers of the invention include the nucleotide sequences B1 to B38 and C1 to C38, preferably B1 to B15, B22, B24, B25, B27 to 29, B32, B34 to B38, C1 to C15, C22, C24, C25, C27 to 29, C32, and C34 to C38, detailed further in Example 2, Table 1.

**Sequencing Of Amplified Genomic DNA And Identification Of Single Nucleotide Polymorphisms**

The amplification products generated as described above, are then sequenced using any method known and available to the skilled technician. Methods for sequencing DNA using either the dideoxy-mediated method (Sanger method) or the Maxam-Gilbert method are widely known to those of ordinary skill in the art. Such methods are for example disclosed in Sambrook et al.(1989). Alternative approaches include hybridization to high-density DNA probe arrays as described in Chee et al.(1996).

Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. The products of the sequencing reactions are run on sequencing gels and the sequences are determined using gel image analysis. The polymorphism search is based on the presence of superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position. Because each dideoxy terminator is labeled with a different fluorescent molecule, the two peaks corresponding to a biallelic site present distinct colors corresponding to two different nucleotides at the same position on the sequence. However, the presence of two peaks can be an artifact due to background noise. To exclude such an artifact, the two DNA strands are sequenced and a comparison between the peaks is carried out. In order to be registered as a polymorphic sequence, the polymorphism has to be detected on both strands.

The above procedure permits those amplification products, which contain biallelic markers to be identified. The detection limit for the frequency of biallelic polymorphisms detected by sequencing pools of 100 individuals is approximately 0.1 for the minor allele, as verified by sequencing pools of known allelic frequencies. However, more than 90% of the biallelic polymorphisms detected by the pooling method have a frequency for the minor allele higher than 0.25. Therefore, the biallelic markers selected by this method have a frequency of at least 0.1 for the minor allele and less than 0.9 for the major allele. Preferably at least 0.2 for the minor allele and less than 0.8 for the major allele, more preferably at least 0.3 for the minor allele and less than 0.7 for the

major allele, thus a heterozygosity rate higher than 0.18, preferably higher than 0.32, more preferably higher than 0.42.

In another embodiment, biallelic markers are detected by sequencing individual DNA samples, the frequency of the minor allele of such a biallelic marker may be less than 0.1.

**Validation Of The Biallelic Markers Of The Present Invention**

The polymorphisms are evaluated for their usefulness as genetic markers by validating that both alleles are present in a population. Validation of the biallelic markers is accomplished by genotyping a group of individuals by a method of the invention and demonstrating that both alleles are present. Microsequencing is a preferred method of genotyping alleles. The validation by genotyping step may be performed on individual samples derived from each individual in the group or by genotyping a pooled sample derived from more than one individual. The group can be as small as one individual if that individual is heterozygous for the allele in question. Preferably the group contains at least three individuals, more preferably the group contains five or six individuals, so that a single validation test will be more likely to result in the validation of more of the biallelic markers that are being tested. It should be noted, however, that when the validation test is performed on a small group it may result in a false negative result if as a result of sampling error none of the individuals tested carries one of the two alleles. Thus, the validation process is less useful in demonstrating that a particular initial result is an artifact, than it is at demonstrating that there is a *bona fide* biallelic marker at a particular position in a sequence. All of the genotyping, haplotyping, association, and interaction study methods of the invention may optionally be performed solely with validated biallelic markers.

**Evaluation Of The Frequency Of The Biallelic Markers Of The Present Invention**

The validated biallelic markers are further evaluated for their usefulness as genetic markers by determining the frequency of the least common allele at the biallelic marker site. The higher the frequency of the less common allele the greater the usefulness of the biallelic marker is association and interaction studies. The determination of the least common allele is accomplished by genotyping a group of individuals by a method of the invention and demonstrating that both alleles are present. This determination of frequency by genotyping step may be performed on individual samples derived from each individual in the group or by genotyping a pooled sample derived from more than one individual. The group must be large enough to be representative of the population as a whole. Preferably the group contains at least 20 individuals, more preferably the group contains at least 50 individuals, most preferably the group contains at least 100 individuals. Of course the larger the group the greater the accuracy of the frequency determination because of reduced sampling error. A biallelic marker wherein the frequency of the less common allele is 30% or more is termed a "high quality biallelic marker." All of the genotyping, haplotyping, association, and interaction study methods of the invention may optionally be performed solely with high quality biallelic markers.

**Methods For Genotyping An Individual For Biallelic Markers**

Methods are provided to genotype a biological sample for one or more biallelic markers of the present invention, all of which may be performed *in vitro*. Such methods of genotyping comprise determining the identity of a nucleotide at a *BAP28* biallelic marker site by any method known in the art. These methods find use in genotyping case-control populations in association studies as well as individuals in the context of detection of alleles of biallelic markers which are known to be associated with a given trait, in which case both copies of the biallelic marker present in individual's genome are determined so that an individual may be classified as homozygous or heterozygous for a particular allele.

These genotyping methods can be performed on nucleic acid samples derived from a single individual or pooled DNA samples.

Genotyping can be performed using similar methods as those described above for the identification of the biallelic markers, or using other genotyping methods such as those further described below. In preferred embodiments, the comparison of sequences of amplified genomic fragments from different individuals is used to identify new biallelic markers whereas microsequencing is used for genotyping known biallelic markers in diagnostic and association study applications.

In one embodiment the invention encompasses methods of genotyping comprising determining the identity of a nucleotide at a *BAP28*-related biallelic marker or the complement thereof in a biological sample; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, wherein said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said biological sample is derived from a single subject; In some embodiments, the identity of the nucleotides at said biallelic marker is determined for both copies of said biallelic marker present in said individual's genome; In some embodiments, said biological sample is derived from multiple subjects; In some embodiments, the method further comprises amplifying a portion of said sequence comprising the biallelic marker prior to said determining step; In some embodiments, said amplifying is performed by PCR; In some embodiments, said determining is performed by a hybridization assay, a sequencing assay, a microsequencing assay, or an enzyme-based mismatch detection assay.

**Source of DNA for genotyping**

Any source of nucleic acids, in purified or non-purified form, can be utilized as the starting nucleic acid, provided it contains or is suspected of containing the specific nucleic acid sequence desired. DNA or RNA may be extracted from cells, tissues, body fluids and the like as described above. While nucleic acids for use in the genotyping methods of the invention can be derived from any mammalian source, the test subjects and individuals from which nucleic acid samples are taken are generally understood to be human.

**Amplification Of DNA Fragments Comprising Biallelic Markers**

Methods and polynucleotides are provided to amplify a segment of nucleotides comprising one or more biallelic marker of the present invention. It will be appreciated that amplification of DNA fragments comprising biallelic markers may be used in various methods and for various purposes and is not restricted to genotyping. Nevertheless, many genotyping methods, although not all, require the previous amplification of the DNA region carrying the biallelic marker of interest. Such methods specifically increase the concentration or total number of sequences that span the biallelic marker or include that site and sequences located either distal or proximal to it. Diagnostic assays may also rely on amplification of DNA segments carrying a biallelic marker of the present invention. Amplification of DNA may be achieved by any method known in the art. Amplification techniques are described above in the section entitled, "Amplification of the *BAP28* gene".

Some of these amplification methods are particularly suited for the detection of single nucleotide polymorphisms and allow the simultaneous amplification of a target sequence and the identification of the polymorphic nucleotide as it is further described below.

The identification of biallelic markers as described above allows the design of appropriate oligonucleotides, which can be used as primers to amplify DNA fragments comprising the biallelic markers of the present invention.

In some embodiments the present invention provides primers for amplifying a DNA fragment containing one or more biallelic markers of the present invention.

The spacing of the primers determines the length of the segment to be amplified. In the context of the present invention, amplified segments carrying biallelic markers can range in size from at least about 25 bp to 35 kbp. Amplification fragments from 25-3000 bp are typical, fragments from 50-1000 bp are preferred and fragments from 100-600 bp are highly preferred. It will be appreciated that amplification primers for the biallelic markers may be any sequence which allow the specific amplification of any DNA fragment carrying the markers. Amplification primers may be labeled or immobilized on a solid support as described in "Oligonucleotide probes and primers".

**Methods of Genotyping DNA samples for Biallelic Markers**

Any method known in the art can be used to identify the nucleotide present at a biallelic marker site. Since the biallelic marker allele to be detected has been identified and specified in the present invention, detection will prove simple for one of ordinary skill in the art by employing any of a number of techniques. Many genotyping methods require the previous amplification of the DNA region carrying the biallelic marker of interest. While the amplification of target or signal is often preferred at present, ultrasensitive detection methods which do not require amplification are also encompassed by the present genotyping methods. Methods well-known to those skilled in the art that can be used to detect biallelic polymorphisms include methods such as, conventional dot blot analyzes, single strand conformational polymorphism analysis (SSCP) described by Orita et al.(1989), denaturing gradient gel electrophoresis (DGGE), heteroduplex analysis, mismatch cleavage detection, and other conventional techniques as described in Sheffield et al.(1991), White et al.(1992), Grompe et al.(1989 and 1993). Another method for determining the identity of the nucleotide present at a particular polymorphic site employs a specialized exonuclease-resistant nucleotide derivative as described in US patent 4,656,127.

Preferred methods involve directly determining the identity of the nucleotide present at a biallelic marker site by sequencing assay, enzyme-based mismatch detection assay, or hybridization assay. The following is a description of some preferred methods. A highly preferred method is the microsequencing technique. The term "sequencing" is used herein to refer to polymerase extension of duplex primer/template complexes and includes both traditional sequencing and microsequencing.

1) Sequencing Assays

The nucleotide present at a polymorphic site can be determined by sequencing methods. In a preferred embodiment, DNA samples are subjected to PCR amplification before sequencing as described above. DNA sequencing methods are described in "Sequencing Of Amplified Genomic DNA And Identification Of Single Nucleotide Polymorphisms".

Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. Sequence analysis allows the identification of the base present at the biallelic marker site.

2) Microsequencing Assays

In microsequencing methods, the nucleotide at a polymorphic site in a target DNA is detected by a single nucleotide primer extension reaction. This method involves appropriate microsequencing primers which, hybridize just upstream of the polymorphic base of interest in the target nucleic acid. A polymerase is used to specifically extend the 3' end of the primer with one single ddNTP (chain terminator) complementary to the nucleotide at the polymorphic site. Next the identity of the incorporated nucleotide is determined in any suitable way.

Typically, microsequencing reactions are carried out using fluorescent ddNTPs and the extended microsequencing primers are analyzed by electrophoresis on ABI 377 sequencing

machines to determine the identity of the incorporated nucleotide as described in EP 412 883. Alternatively capillary electrophoresis can be used in order to process a higher number of assays simultaneously. An example of a typical microsequencing procedure that can be used in the context of the present invention is provided in Example 4.

Different approaches can be used for the labeling and detection of ddNTPs. A homogeneous phase detection method based on fluorescence resonance energy transfer has been described by Chen and Kwok (1997) and Chen et al.(1997). In this method, amplified genomic DNA fragments containing polymorphic sites are incubated with a 5`-fluorescein–labeled primer in the presence of allelic dye-labeled dideoxyribonucleoside triphosphates and a modified Taq polymerase. The dye-labeled primer is extended one base by the dye-terminator specific for the allele present on the template. At the end of the genotyping reaction, the fluorescence intensities of the two dyes in the reaction mixture are analyzed directly without separation or purification. All these steps can be performed in the same tube and the fluorescence changes can be monitored in real time. Alternatively, the extended primer may be analyzed by MALDI-TOF Mass Spectrometry. The base at the polymorphic site is identified by the mass added onto the microsequencing primer (see Haff and Smirnov, 1997).

Microsequencing may be achieved by the established microsequencing method or by developments or derivatives thereof. Alternative methods include several solid-phase microsequencing techniques. The basic microsequencing protocol is the same as described previously, except that the method is conducted as a heterogeneous phase assay, in which the primer or the target molecule is immobilized or captured onto a solid support. To simplify the primer separation and the terminal nucleotide addition analysis, oligonucleotides are attached to solid supports or are modified in such ways that permit affinity separation as well as polymerase extension. The 5` ends and internal nucleotides of synthetic oligonucleotides can be modified in a number of different ways to permit different affinity separation approaches, e.g., biotinylation. If a single affinity group is used on the oligonucleotides, the oligonucleotides can be separated from the incorporated terminator regent. This eliminates the need of physical or size separation. More than one oligonucleotide can be separated from the terminator reagent and analyzed simultaneously if more than one affinity group is used. This permits the analysis of several nucleic acid species or more nucleic acid sequence information per extension reaction. The affinity group need not be on the priming oligonucleotide but could alternatively be present on the template. For example, immobilization can be carried out via an interaction between biotinylated DNA and streptavidin-coated microtitration wells or avidin-coated polystyrene particles. In the same manner, oligonucleotides or templates may be attached to a solid support in a high-density format. In such solid phase microsequencing reactions, incorporated ddNTPs can be radiolabeled (Syvänen, 1994) or linked to fluorescein (Livak and Hainer, 1994). The detection of radiolabeled ddNTPs can be achieved through scintillation-based techniques. The detection of fluorescein-linked ddNTPs can be

based on the binding of antifluorescein antibody conjugated with alkaline phosphatase, followed by incubation with a chromogenic substrate (such as *p*-nitrophenyl phosphate). Other possible reporter-detection pairs include: ddNTP linked to dinitrophenyl (DNP) and anti-DNP alkaline phosphatase conjugate (Harju et al., 1993) or biotinylated ddNTP and horseradish peroxidase-conjugated streptavidin with *o*-phenylenediamine as a substrate (WO 92/15712). As yet another alternative solid-phase microsequencing procedure, Nyren et al.(1993) described a method relying on the detection of DNA polymerase activity by an enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA).

Pastinen et al.(1997) describe a method for multiplex detection of single nucleotide polymorphism in which the solid phase minisequencing principle is applied to an oligonucleotide array format. High-density arrays of DNA probes attached to a solid support (DNA chips) are further described below.

In one aspect the present invention provides polynucleotides and methods to genotype one or more biallelic markers of the present invention by performing a microsequencing assay. Preferred microsequencing primers include the nucleotide sequences D1 to D58 and E1 to E58, preferably D1 to D27, D34, D37 to D41, D43 to D49, D52, D54 to D58, E1 to E27, E34, E37 to E41, E43 to E49, E52, and E54 to E58. It will be appreciated that the microsequencing primers listed in Example 4 are merely exemplary and that, any primer having a 3' end immediately adjacent to the polymorphic nucleotide may be used. Similarly, it will be appreciated that microsequencing analysis may be performed for any biallelic marker or any combination of biallelic markers of the present invention. One aspect of the present invention is a solid support which includes one or more microsequencing primers listed in Example 4, or fragments comprising at least 8, 12, 15, 20, 25, 30, 40, or 50 consecutive nucleotides thereof and having a 3' terminus immediately upstream of the corresponding biallelic marker, for determining the identity of a nucleotide at a biallelic marker site.

3) Mismatch detection assays based on polymerases and ligases

In one aspect the present invention provides polynucleotides and methods to determine the allele of one or more biallelic markers of the present invention in a biological sample, by mismatch detection assays based on polymerases and/or ligases. These assays are based on the specificity of polymerases and ligases. Polymerization reactions places particularly stringent requirements on correct base pairing of the 3' end of the amplification primer and the joining of two oligonucleotides hybridized to a target DNA sequence is quite sensitive to mismatches close to the ligation site, especially at the 3' end. Methods, primers and various parameters to amplify DNA fragments comprising biallelic markers of the present invention are further described above in "Amplification Of DNA Fragments Comprising Biallelic Markers".

**Allele Specific Amplification Primers**

Discrimination between the two alleles of a biallelic marker can also be achieved by allele specific amplification, a selective strategy, whereby one of the alleles is amplified without

amplification of the other allele. This is accomplished by placing the polymorphic base at the 3' end of one of the amplification primers. Because the extension forms from the 3'end of the primer, a mismatch at or near this position has an inhibitory effect on amplification. Therefore, under appropriate amplification conditions, these primers only direct amplification on their complementary allele. Determining the precise location of the mismatch and the corresponding assay conditions are well with the ordinary skill in the art.

**Ligation/Amplification Based Methods**

The "Oligonucleotide Ligation Assay" (OLA) uses two oligonucleotides which are designed to be capable of hybridizing to abutting sequences of a single strand of a target molecules. One of the oligonucleotides is biotinylated, and the other is detectably labeled. If the precise complementary sequence is found in a target molecule, the oligonucleotides will hybridize such that their termini abut, and create a ligation substrate that can be captured and detected. OLA is capable of detecting single nucleotide polymorphisms and may be advantageously combined with PCR as described by Nickerson et al.(1990). In this method, PCR is used to achieve the exponential amplification of target DNA, which is then detected using OLA.

Other amplification methods which are particularly suited for the detection of single nucleotide polymorphism include LCR (ligase chain reaction), Gap LCR (GLCR) which are described above in "Amplification of the *BAP28* gene". LCR uses two pairs of probes to exponentially amplify a specific target. The sequences of each pair of oligonucleotides, is selected to permit the pair to hybridize to abutting sequences of the same strand of the target. Such hybridization forms a substrate for a template-dependant ligase. In accordance with the present invention, LCR can be performed with oligonucleotides having the proximal and distal sequences of the same strand of a biallelic marker site. In one embodiment, either oligonucleotide will be designed to include the biallelic marker site. In such an embodiment, the reaction conditions are selected such that the oligonucleotides can be ligated together only if the target molecule either contains or lacks the specific nucleotide that is complementary to the biallelic marker on the oligonucleotide. In an alternative embodiment, the oligonucleotides will not include the biallelic marker, such that when they hybridize to the target molecule, a "gap" is created as described in WO 90/01069. This gap is then "filled" with complementary dNTPs (as mediated by DNA polymerase), or by an additional pair of oligonucleotides. Thus at the end of each cycle, each single strand has a complement capable of serving as a target during the next cycle and exponential allele-specific amplification of the desired sequence is obtained.

Ligase/Polymerase-mediated Genetic Bit Analysis™ is another method for determining the identity of a nucleotide at a preselected site in a nucleic acid molecule (WO 95/21271). This method involves the incorporation of a nucleoside triphosphate that is complementary to the nucleotide present at the preselected site onto the terminus of a primer molecule, and their subsequent ligation

to a second oligonucleotide. The reaction is monitored by detecting a specific label attached to the reaction's solid phase or by detection in solution.

4) Hybridization Assay Methods

A preferred method of determining the identity of the nucleotide present at a biallelic marker site involves nucleic acid hybridization. The hybridization probes, which can be conveniently used in such reactions, preferably include the probes defined herein. Any hybridization assay may be used including Southern hybridization, Northern hybridization, dot blot hybridization and solid-phase hybridization (see Sambrook et al., 1989).

Hybridization refers to the formation of a duplex structure by two single stranded nucleic acids due to complementary base pairing. Hybridization can occur between exactly complementary nucleic acid strands or between nucleic acid strands that contain minor regions of mismatch. Specific probes can be designed that hybridize to one form of a biallelic marker and not to the other and therefore are able to discriminate between different allelic forms. Allele-specific probes are often used in pairs, one member of a pair showing perfect match to a target sequence containing the original allele and the other showing a perfect match to the target sequence containing the alternative allele. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Stringent, sequence specific hybridization conditions, under which a probe will hybridize only to the exactly complementary target sequence are well known in the art (Sambrook et al., 1989). Stringent conditions are sequence dependent and will be different in different circumstances. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. Although such hybridizations can be performed in solution, it is preferred to employ a solid-phase hybridization assay. The target DNA comprising a biallelic marker of the present invention may be amplified prior to the hybridization reaction. The presence of a specific allele in the sample is determined by detecting the presence or the absence of stable hybrid duplexes formed between the probe and the target DNA. The detection of hybrid duplexes can be carried out by a number of methods. Various detection assay formats are well known which utilize detectable labels bound to either the target or the probe to enable detection of the hybrid duplexes. Typically, hybridization duplexes are separated from unhybridized nucleic acids and the labels bound to the duplexes are then detected. Those skilled in the art will recognize that wash steps may be employed to wash away excess target DNA or probe as well as unbound conjugate. Further, standard heterogeneous assay formats are suitable for detecting the hybrids using the labels present on the primers and probes.

Two recently developed assays allow hybridization-based allele discrimination with no need for separations or washes (see Landegren U. et al., 1998). The TaqMan assay takes advantage of the 5' nuclease activity of Taq DNA polymerase to digest a DNA probe annealed specifically to the accumulating amplification product. TaqMan probes are labeled with a donor-acceptor dye pair

that interacts via fluorescence energy transfer. Cleavage of the TaqMan probe by the advancing polymerase during amplification dissociates the donor dye from the quenching acceptor dye, greatly increasing the donor fluorescence. All reagents necessary to detect two allelic variants can be assembled at the beginning of the reaction and the results are monitored in real time (see Livak et al., 1995). In an alternative homogeneous hybridization based procedure, molecular beacons are used for allele discriminations. Molecular beacons are hairpin-shaped oligonucleotide probes that report the presence of specific nucleic acids in homogeneous solutions. When they bind to their targets they undergo a conformational reorganization that restores the fluorescence of an internally quenched fluorophore (Tyagi et al., 1998).

The polynucleotides provided herein can be used to produce probes which can be used in hybridization assays for the detection of biallelic marker alleles in biological samples. These probes are characterized in that they preferably comprise between 8 and 50 nucleotides, and in that they are sufficiently complementary to a sequence comprising a biallelic marker of the present invention to hybridize thereto and preferably sufficiently specific to be able to discriminate the targeted sequence for only one nucleotide variation. A particularly preferred probe is 25 nucleotides in length. Preferably the biallelic marker is within 4 nucleotides of the center of the polynucleotide probe. In particularly preferred probes, the biallelic marker is at the center of said polynucleotide. Preferred probes comprise a nucleotide sequence selected from the group consisting of amplicons listed in Table 1 and the sequences complementary thereto, or a fragment thereof, said fragment comprising at least about 8 consecutive nucleotides, preferably 10, 15, 20, more preferably 25, 30, 40, 47, or 50 consecutive nucleotides and containing a polymorphic base. Preferred probes comprise a nucleotide sequence selected from the group consisting of P1 to P58 and the sequences complementary thereto, preferably P1 to P27, P34, P37 to P41, P43 to P49, P52, P54 to P58. In preferred embodiments the polymorphic base(s) are within 5, 4, 3, 2, 1, nucleotides of the center of the said polynucleotide, more preferably at the center of said polynucleotide.

Preferably the probes of the present invention are labeled or immobilized on a solid support. Labels and solid supports are further described in "Oligonucleotide Probes and Primers". The probes can be non-extendable as described in "Oligonucleotide Probes and Primers".

By assaying the hybridization to an allele specific probe, one can detect the presence or absence of a biallelic marker allele in a given sample. High-Throughput parallel hybridizations in array format are specifically encompassed within "hybridization assays" and are described below.

5) Hybridization To Addressable Arrays Of Oligonucleotides

Hybridization assays based on oligonucleotide arrays rely on the differences in hybridization stability of short oligonucleotides to perfectly matched and mismatched target sequence variants. Efficient access to polymorphism information is obtained through a basic structure comprising high-density arrays of oligonucleotide probes attached to a solid support (e.g.,

the chip) at selected positions. Each DNA chip can contain thousands to millions of individual synthetic DNA probes arranged in a grid-like pattern and miniaturized to the size of a dime.

The chip technology has already been applied with success in numerous cases. For example, the screening of mutations has been undertaken in the BRCA1 gene, in *S. cerevisiae* mutant strains, and in the protease gene of HIV-1 virus (Hacia et al., 1996; Shoemaker et al., 1996; Kozal et al., 1996). Chips of various formats for use in detecting biallelic polymorphisms can be produced on a customized basis by Affymetrix (GeneChip™), Hyseq (HyChip and HyGnostics), and Protogene Laboratories.

In general, these methods employ arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from an individual which, target sequences include a polymorphic marker. EP 785280 describes a tiling strategy for the detection of single nucleotide polymorphisms. Briefly, arrays may generally be "tiled" for a large number of specific polymorphisms. By "tiling" is generally meant the synthesis of a defined set of oligonucleotide probes which is made up of a sequence complementary to the target sequence of interest, as well as preselected variations of that sequence, e.g., substitution of one or more given positions with one or more members of the basis set of monomers, i.e. nucleotides. Tiling strategies are further described in PCT application No WO 95/11995. In a particular aspect, arrays are tiled for a number of specific, identified biallelic marker sequences. In particular, the array is tiled to include a number of detection blocks, each detection block being specific for a specific biallelic marker or a set of biallelic markers. For example, a detection block may be tiled to include a number of probes, which span the sequence segment that includes a specific polymorphism. To ensure probes that are complementary to each allele, the probes are synthesized in pairs differing at the biallelic marker. In addition to the probes differing at the polymorphic base, monosubstituted probes are also generally tiled within the detection block. These monosubstituted probes have bases at and up to a certain number of bases in either direction from the polymorphism, substituted with the remaining nucleotides (selected from A, T, G, C and U). Typically the probes in a tiled detection block will include substitutions of the sequence positions up to and including those that are 5 bases away from the biallelic marker. The monosubstituted probes provide internal controls for the tiled array, to distinguish actual hybridization from artefactual cross-hybridization. Upon completion of hybridization with the target sequence and washing of the array, the array is scanned to determine the position on the array to which the target sequence hybridizes. The hybridization data from the scanned array is then analyzed to identify which allele or alleles of the biallelic marker are present in the sample. Hybridization and scanning may be carried out as described in PCT application No WO 92/10092 and WO 95/11995 and US patent No 5,424,186.

Thus, in some embodiments, the chips may comprise an array of nucleic acid sequences of fragments of about 15 nucleotides in length. In further embodiments, the chip may comprise an array including at least one sequences comprising at least about 8 consecutive nucleotides,

preferably 10, 15, 20, more preferably 25, 30, 40, 47, or 50 consecutive nucleotides and containing a polymorphic base. In preferred embodiments the polymorphic base is within 5, 4, 3, 2, 1, nucleotides of the center of the said polynucleotide, more preferably at the center of said polynucleotide. In some embodiments, the chip may comprise an array of at least 2, 3, 4, 5, 6, 7, 8 or more of these polynucleotides of the invention. Solid supports and polynucleotides of the present invention attached to solid supports are further described in "oligonucleotide probes and primers".

6) Integrated Systems

Another technique, which may be used to analyze polymorphisms, includes multicomponent integrated systems, which miniaturize and compartmentalize processes such as PCR and capillary electrophoresis reactions in a single functional device. An example of such technique is disclosed in US patent 5,589,136, which describes the integration of PCR amplification and capillary electrophoresis in chips.

Integrated systems can be envisaged mainly when microfluidic systems are used. These systems comprise a pattern of microchannels designed onto a glass, silicon, quartz, or plastic wafer included on a microchip. The movements of the samples are controlled by electric, electroosmotic or hydrostatic forces applied across different areas of the microchip to create functional microscopic valves and pumps with no moving parts.

For genotyping biallelic markers, the microfluidic system may integrate nucleic acid amplification, microsequencing, capillary electrophoresis and a detection method such as laser-induced fluorescence detection.

**Methods Of Genetic Analysis Using The Biallelic Markers Of The Present Invention**

Different methods are available for the genetic analysis of complex traits (see Lander and Schork, 1994). The search for disease-susceptibility genes is conducted using two main methods: the linkage approach in which evidence is sought for cosegregation between a locus and a putative trait locus using family studies, and the association approach in which evidence is sought for a statistically significant association between an allele and a trait or a trait causing allele (Khoury et al., 1993). In general, the biallelic markers of the present invention find use in any method known in the art to demonstrate a statistically significant correlation between a genotype and a phenotype. The biallelic markers may be used in parametric and non-parametric linkage analysis methods. Preferably, the biallelic markers of the present invention are used to identify genes associated with detectable traits using association studies, an approach which does not require the use of affected families and which permits the identification of genes associated with complex and sporadic traits.

The genetic analysis using the biallelic markers of the present invention may be conducted on any scale. The whole set of biallelic markers of the present invention or any subset of biallelic markers of the present invention corresponding to the candidate gene may be used. Further, any set of genetic markers including a biallelic marker of the present invention may be used. A set of biallelic polymorphisms that could be used as genetic markers in combination with the biallelic

markers of the present invention has been described in WO 98/20165. As mentioned above, it should be noted that the biallelic markers of the present invention may be included in any complete or partial genetic map of the human genome. These different uses are specifically contemplated in the present invention and claims.

**Linkage Analysis**

Linkage analysis is based upon establishing a correlation between the transmission of genetic markers and that of a specific trait throughout generations within a family. Thus, the aim of linkage analysis is to detect marker loci that show cosegregation with a trait of interest in pedigrees.

Parametric Methods

When data are available from successive generations there is the opportunity to study the degree of linkage between pairs of loci. Estimates of the recombination fraction enable loci to be ordered and placed onto a genetic map. With loci that are genetic markers, a genetic map can be established, and then the strength of linkage between markers and traits can be calculated and used to indicate the relative positions of markers and genes affecting those traits (Weir, 1996). The classical method for linkage analysis is the logarithm of odds (lod) score method (see Morton, 1955; Ott, 1991). Calculation of lod scores requires specification of the mode of inheritance for the disease (parametric method). Generally, the length of the candidate region identified using linkage analysis is between 2 and 20Mb. Once a candidate region is identified as described above, analysis of recombinant individuals using additional markers allows further delineation of the candidate region. Linkage analysis studies have generally relied on the use of a maximum of 5,000 microsatellite markers, thus limiting the maximum theoretical attainable resolution of linkage analysis to about 600 kb on average.

Linkage analysis has been successfully applied to map simple genetic traits that show clear Mendelian inheritance patterns and which have a high penetrance (i.e., the ratio between the number of trait positive carriers of allele *a* and the total number of *a* carriers in the population). However, parametric linkage analysis suffers from a variety of drawbacks. First, it is limited by its reliance on the choice of a genetic model suitable for each studied trait. Furthermore, as already mentioned, the resolution attainable using linkage analysis is limited, and complementary studies are required to refine the analysis of the typical 2Mb to 20Mb regions initially identified through linkage analysis. In addition, parametric linkage analysis approaches have proven difficult when applied to complex genetic traits, such as those due to the combined action of multiple genes and/or environmental factors. It is very difficult to model these factors adequately in a lod score analysis. In such cases, too large an effort and cost are needed to recruit the adequate number of affected families required for applying linkage analysis to these situations, as recently discussed by Risch, N. and Merikangas, K. (1996).

Non-Parametric Methods

The advantage of the so-called non-parametric methods for linkage analysis is that they do not require specification of the mode of inheritance for the disease, they tend to be more useful for the analysis of complex traits. In non-parametric methods, one tries to prove that the inheritance pattern of a chromosomal region is not consistent with random Mendelian segregation by showing that affected relatives inherit identical copies of the region more often than expected by chance. Affected relatives should show excess "allele sharing" even in the presence of incomplete penetrance and polygenic inheritance. In non-parametric linkage analysis the degree of agreement at a marker locus in two individuals can be measured either by the number of alleles identical by state (IBS) or by the number of alleles identical by descent (IBD). Affected sib pair analysis is a well-known special case and is the simplest form of these methods.

The biallelic markers of the present invention may be used in both parametric and non-parametric linkage analysis. Preferably biallelic markers may be used in non-parametric methods which allow the mapping of genes involved in complex traits. The biallelic markers of the present invention may be used in both IBD- and IBS- methods to map genes affecting a complex trait. In such studies, taking advantage of the high density of biallelic markers, several adjacent biallelic marker loci may be pooled to achieve the efficiency attained by multi-allelic markers (Zhao et al., 1998).

**Population Association Studies**

The present invention comprises methods for identifying if the *BAP28* gene is associated with a detectable trait using the biallelic markers of the present invention. In one embodiment the present invention comprises methods to detect an association between a biallelic marker allele or a biallelic marker haplotype and a trait. Further, the invention comprises methods to identify a trait causing allele in linkage disequilibrium with any biallelic marker allele of the present invention.

As described above, alternative approaches can be employed to perform association studies: genome-wide association studies, candidate region association studies and candidate gene association studies. In a preferred embodiment, the biallelic markers of the present invention are used to perform candidate gene association studies. The candidate gene analysis clearly provides a short-cut approach to the identification of genes and gene polymorphisms related to a particular trait when some information concerning the biology of the trait is available. Further, the biallelic markers of the present invention may be incorporated in any map of genetic markers of the human genome in order to perform genome-wide association studies. Methods to generate a high-density map of biallelic markers has been described in US Provisional Patent application serial number 60/082,614. The biallelic markers of the present invention may further be incorporated in any map of a specific candidate region of the genome (a specific chromosome or a specific chromosomal segment for example).

As mentioned above, association studies may be conducted within the general population and are not limited to studies performed on related individuals in affected families. Association studies are extremely valuable as they permit the analysis of sporadic or multifactor traits. Moreover, association studies represent a powerful method for fine-scale mapping enabling much finer mapping of trait causing alleles than linkage studies. Studies based on pedigrees often only narrow the location of the trait causing allele. Association studies using the biallelic markers of the present invention can therefore be used to refine the location of a trait causing allele in a candidate region identified by Linkage Analysis methods. Moreover, once a chromosome segment of interest has been identified, the presence of a candidate gene such as a candidate gene of the present invention, in the region of interest can provide a shortcut to the identification of the trait causing allele. Biallelic markers of the present invention can be used to demonstrate that a candidate gene is associated with a trait. Such uses are specifically contemplated in the present invention.

**Determining The Frequency Of A Biallelic Marker Allele Or Of A Biallelic Marker Haplotype In A Population**

Association studies explore the relationships among frequencies for sets of alleles between loci.

Determining The Frequency Of An Allele In A Population

Allelic frequencies of the biallelic markers in a populations can be determined using one of the methods described above under the heading "Methods for genotyping an individual for biallelic markers", or any genotyping procedure suitable for this intended purpose. Genotyping pooled samples or individual samples can determine the frequency of a biallelic marker allele in a population. One way to reduce the number of genotypings required is to use pooled samples. A major obstacle in using pooled samples is in terms of accuracy and reproducibility for determining accurate DNA concentrations in setting up the pools. Genotyping individual samples provides higher sensitivity, reproducibility and accuracy and; is the preferred method used in the present invention. Preferably, each individual is genotyped separately and simple gene counting is applied to determine the frequency of an allele of a biallelic marker or of a genotype in a given population.

The invention also relates to methods of estimating the frequency of an allele in a population comprising: a) genotyping individuals from said population for said biallelic marker according to the method of the present invention; b) determining the proportional representation of said biallelic marker in said population. In addition, the methods of estimating the frequency of an allele in a population of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to

A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, the step of determining the frequency of a biallelic marker allele in a population may be accomplished by determining the identity of the nucleotides for both copies of said biallelic marker present in the genome of each individual in said population and calculating the proportional representation of said nucleotide at said *BAP28*-related biallelic marker for the population; In some embodiments, the step of determining the proportional representation may be accomplished by performing a genotyping method of the invention on a pooled biological sample derived from a representative number of individuals, or each individual, in said population, and calculating the proportional amount of said nucleotide compared with the total.

Determining The Frequency Of A Haplotype In A Population

The gametic phase of haplotypes is unknown when diploid individuals are heterozygous at more than one locus. Using genealogical information in families gametic phase can sometimes be inferred (Perlin et al., 1994). When no genealogical information is available different strategies may be used. One possibility is that the multiple-site heterozygous diploids can be eliminated from the analysis, keeping only the homozygotes and the single-site heterozygote individuals, but this approach might lead to a possible bias in the sample composition and the underestimation of low-frequency haplotypes. Another possibility is that single chromosomes can be studied independently, for example, by asymmetric PCR amplification (see Newton et al. 1989; Wu et al., 1989) or by isolation of single chromosome by limit dilution followed by PCR amplification (see Ruano et al., 1990). Further, a sample may be haplotyped for sufficiently close biallelic markers by double PCR amplification of specific alleles (Sarkar, G. and Sommer S. S., 1991). These approaches are not entirely satisfying either because of their technical complexity, the additional cost they entail, their lack of generalization at a large scale, or the possible biases they introduce. To overcome these difficulties, an algorithm to infer the phase of PCR-amplified DNA genotypes introduced by Clark, A.G.(1990) may be used. Briefly, the principle is to start filling a preliminary list of haplotypes present in the sample by examining unambiguous individuals, that is, the complete homozygotes and the single-site heterozygotes. Then other individuals in the same sample are screened for the possible occurrence of previously recognized haplotypes. For each positive identification, the complementary haplotype is added to the list of recognized haplotypes, until the phase information for all individuals is either resolved or identified as unresolved. This method assigns a single haplotype to each multiheterozygous individual, whereas several haplotypes are possible when there are more than one heterozygous site. Alternatively, one can use methods estimating haplotype frequencies in a population without assigning haplotypes to each individual. Preferably, a method based on an expectation-maximization (EM) algorithm (Dempster et al., 1977) leading to maximum-

likelihood estimates of haplotype frequencies under the assumption of Hardy-Weinberg proportions (random mating) is used (see Excoffier L. and Slatkin M., 1995). The EM algorithm is a generalized iterative maximum-likelihood approach to estimation that is useful when data are ambiguous and/or incomplete. The EM algorithm is used to resolve heterozygotes into haplotypes. Haplotype estimations are further described below under the heading "Statistical Methods." Any other method known in the art to determine or to estimate the frequency of a haplotype in a population may be used.

The invention also encompasses methods of estimating the frequency of a haplotype for a set of biallelic markers in a population, comprising the steps of: a) genotyping at least one *BAP28*-related biallelic marker according to a method of the invention for each individual in said population; b) genotyping a second biallelic marker by determining the identity of the nucleotides at said second biallelic marker for both copies of said second biallelic marker present in the genome of each individual in said population; and c) applying a haplotype determination method to the identities of the nucleotides determined in steps a) and b) to obtain an estimate of said frequency. In addition, the methods of estimating the frequency of a haplotype of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said haplotype determination method is performed by asymmetric PCR amplification, double PCR amplification of specific alleles, the Clark algorithm, or an expectation-maximization algorithm.

**Linkage Disequilibrium Analysis**

Linkage disequilibrium is the non-random association of alleles at two or more loci and represents a powerful tool for mapping genes involved in disease traits (see Ajioka R.S. et al., 1997). Biallelic markers, because they are densely spaced in the human genome and can be genotyped in greater numbers than other types of genetic markers (such as RFLP or VNTR markers), are particularly useful in genetic analysis based on linkage disequilibrium.

When a disease mutation is first introduced into a population (by a new mutation or the immigration of a mutation carrier), it necessarily resides on a single chromosome and thus on a single "background" or "ancestral" haplotype of linked markers. Consequently, there is complete disequilibrium between these markers and the disease mutation: one finds the disease mutation only

in the presence of a specific set of marker alleles. Through subsequent generations recombination events occur between the disease mutation and these marker polymorphisms, and the disequilibrium gradually dissipates. The pace of this dissipation is a function of the recombination frequency, so the markers closest to the disease gene will manifest higher levels of disequilibrium than those that are further away. When not broken up by recombination, "ancestral" haplotypes and linkage disequilibrium between marker alleles at different loci can be tracked not only through pedigrees but also through populations. Linkage disequilibrium is usually seen as an association between one specific allele at one locus and another specific allele at a second locus.

The pattern or curve of disequilibrium between disease and marker loci is expected to exhibit a maximum that occurs at the disease locus. Consequently, the amount of linkage disequilibrium between a disease allele and closely linked genetic markers may yield valuable information regarding the location of the disease gene. For fine-scale mapping of a disease locus, it is useful to have some knowledge of the patterns of linkage disequilibrium that exist between markers in the studied region. As mentioned above the mapping resolution achieved through the analysis of linkage disequilibrium is much higher than that of linkage studies. The high density of biallelic markers combined with linkage disequilibrium analysis provides powerful tools for fine-scale mapping. Different methods to calculate linkage disequilibrium are described below under the heading "Statistical Methods".

**Population-Based Case-Control Studies Of Trait-Marker Associations**

As mentioned above, the occurrence of pairs of specific alleles at different loci on the same chromosome is not random and the deviation from random is called linkage disequilibrium. Association studies focus on population frequencies and rely on the phenomenon of linkage disequilibrium. If a specific allele in a given gene is directly involved in causing a particular trait, its frequency will be statistically increased in an affected (trait positive) population, when compared to the frequency in a trait negative population or in a random control population. As a consequence of the existence of linkage disequilibrium, the frequency of all other alleles present in the haplotype carrying the trait-causing allele will also be increased in trait positive individuals compared to trait negative individuals or random controls. Therefore, association between the trait and any allele (specifically a biallelic marker allele) in linkage disequilibrium with the trait-causing allele will suffice to suggest the presence of a trait-related gene in that particular region. Case-control populations can be genotyped for biallelic markers to identify associations that narrowly locate a trait causing allele. As any marker in linkage disequilibrium with one given marker associated with a trait will be associated with the trait. Linkage disequilibrium allows the relative frequencies in case-control populations of a limited number of genetic polymorphisms (specifically biallelic markers) to be analyzed as an alternative to screening all possible functional polymorphisms in order

to find trait-causing alleles. Association studies compare the frequency of marker alleles in unrelated case-control populations, and represent powerful tools for the dissection of complex traits.

Case-Control Populations (Inclusion Criteria)

Population-based association studies do not concern familial inheritance but compare the prevalence of a particular genetic marker, or a set of markers, in case-control populations. They are case-control studies based on comparison of unrelated case (affected or trait positive) individuals and unrelated control (unaffected, trait negative or random) individuals. Preferably the control group is composed of unaffected or trait negative individuals. Further, the control group is ethnically matched to the case population. Moreover, the control group is preferably matched to the case-population for the main known confusion factor for the trait under study (for example age-matched for an age-dependent trait). Ideally, individuals in the two samples are paired in such a way that they are expected to differ only in their disease status. The terms "trait positive population", "case population" and "affected population" are used interchangeably herein.

An important step in the dissection of complex traits using association studies is the choice of case-control populations (see Lander and Schork, 1994). A major step in the choice of case-control populations is the clinical definition of a given trait or phenotype. Any genetic trait may be analyzed by the association method proposed here by carefully selecting the individuals to be included in the trait positive and trait negative phenotypic groups. Four criteria are often useful: clinical phenotype, age at onset, family history and severity. The selection procedure for continuous or quantitative traits (such as blood pressure for example) involves selecting individuals at opposite ends of the phenotype distribution of the trait under study, so as to include in these trait positive and trait negative populations individuals with non-overlapping phenotypes. Preferably, case-control populations are phenotypically homogeneous populations. Trait positive and trait negative populations consist of phenotypically uniform populations of individuals representing each between 1 and 98%, preferably between 1 and 80%, more preferably between 1 and 50%, and more preferably between 1 and 30%, most preferably between 1 and 20% of the total population under study, and preferably selected among individuals exhibiting non-overlapping phenotypes. The clearer the difference between the two trait phenotypes, the greater the probability of detecting an association with biallelic markers. The selection of those drastically different but relatively uniform phenotypes enables efficient comparisons in association studies and the possible detection of marked differences at the genetic level, provided that the sample sizes of the populations under study are significant enough.

In preferred embodiments, a first group of between 50 and 300 trait positive individuals, preferably about 100 individuals, are recruited according to their phenotypes. A similar number of control individuals are included in such studies.

In the present invention, typical examples of inclusion criteria include, but are not restricted to, prostate cancer or aggressiveness of prostate cancer tumors. In one preferred

embodiment of the present invention, association studies are carried out on the basis of a presence (trait positive) or absence (trait negative) of prostate cancer.

Associations studies can be carried out by the skilled technician using the biallelic markers of the invention defined above, with different trait positive and trait negative populations. Suitable further examples of association studies using biallelic markers of the *BAP28* gene, including the biallelic markers A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, involve studies on the following populations:

- a trait positive population suffering from a cancer and a healthy unaffected population, or
- a trait positive population suffering from prostate cancer treated with agents acting against prostate cancer and suffering from side-effects resulting from this treatment and an trait negative population suffering from prostate cancer treated with same agents without any substantial side-effects, or
- a trait positive population suffering from prostate cancer treated with agents acting against prostate cancer showing a beneficial response and a trait negative population suffering from prostate cancer treated with same agents without any beneficial response, or
- a trait positive population suffering from prostate cancer presenting highly aggressive prostate cancer tumors and a trait negative population suffering from prostate cancer with prostate cancer tumors devoid of aggressiveness.

Association Analysis

The invention also comprises methods of detecting an association between a genotype and a phenotype, comprising the steps of: a) determining the frequency of at least one *BAP28*-related biallelic marker in a trait positive population according to a genotyping method of the invention; b) determining the frequency of said *BAP28*-related biallelic marker in a control population according to a genotyping method of the invention; and c) determining whether a statistically significant association exists between said genotype and said phenotype. In addition, the methods of detecting an association between a genotype and a phenotype of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said control population may be a trait negative population, or a random population; In some embodiments, each of said genotyping steps a) and b) may be performed on a pooled biological sample derived from each of said populations; In

some embodiments, each of said genotyping of steps a) and b) is performed separately on biological samples derived from each individual in said population or a subsample thereof.

The general strategy to perform association studies using biallelic markers derived from a region carrying a candidate gene is to scan two groups of individuals (case-control populations) in order to measure and statistically compare the allele frequencies of the biallelic markers of the present invention in both groups.

If a statistically significant association with a trait is identified for at least one or more of the analyzed biallelic markers, one can assume that: either the associated allele is directly responsible for causing the trait (i.e. the associated allele is the trait causing allele), or more likely the associated allele is in linkage disequilibrium with the trait causing allele. The specific characteristics of the associated allele with respect to the candidate gene function usually give further insight into the relationship between the associated allele and the trait (causal or in linkage disequilibrium). If the evidence indicates that the associated allele within the candidate gene is most probably not the trait causing allele but is in linkage disequilibrium with the real trait causing allele, then the trait causing allele can be found by sequencing the vicinity of the associated marker, and performing further association studies with the polymorphisms that are revealed in an iterative manner.

Association studies are usually run in two successive steps. In a first phase, the frequencies of a reduced number of biallelic markers from the candidate gene are determined in the trait positive and control populations. In a second phase of the analysis, the position of the genetic loci responsible for the given trait is further refined using a higher density of markers from the relevant region. However, if the candidate gene under study is relatively small in length, as is the case for *BAP28*, a single phase may be sufficient to establish significant associations.

Haplotype Analysis

As described above, when a chromosome carrying a disease allele first appears in a population as a result of either mutation or migration, the mutant allele necessarily resides on a chromosome having a set of linked markers: the ancestral haplotype. This haplotype can be tracked through populations and its statistical association with a given trait can be analyzed. Complementing single point (allelic) association studies with multi-point association studies also called haplotype studies increases the statistical power of association studies. Thus, a haplotype association study allows one to define the frequency and the type of the ancestral carrier haplotype. A haplotype analysis is important in that it increases the statistical power of an analysis involving individual markers.

In a first stage of a haplotype frequency analysis, the frequency of the possible haplotypes based on various combinations of the identified biallelic markers of the invention is determined. The haplotype frequency is then compared for distinct populations of trait positive and control individuals. The number of trait positive individuals, which should be, subjected to this analysis to

obtain statistically significant results usually ranges between 30 and 300, with a preferred number of individuals ranging between 50 and 150. The same considerations apply to the number of unaffected individuals (or random control) used in the study. The results of this first analysis provide haplotype frequencies in case-control populations, for each evaluated haplotype frequency a p-value and an odd ratio are calculated. If a statistically significant association is found the relative risk for an individual carrying the given haplotype of being affected with the trait under study can be approximated.

An additional embodiment of the present invention encompasses methods of detecting an association between a haplotype and a phenotype, comprising the steps of: a) estimating the frequency of at least one haplotype in a trait positive population, according to a method of the invention for estimating the frequency of a haplotype; b) estimating the frequency of said haplotype in a control population, according to a method of the invention for estimating the frequency of a haplotype; and c) determining whether a statistically significant association exists between said haplotype and said phenotype. In addition, the methods of detecting an association between a haplotype and a phenotype of the invention encompass methods with any further limitation described in this disclosure, or those following: In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said *BAP28*-related biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, or the biallelic markers in linkage disequilibrium therewith; In some embodiments, said control population is a trait negative population, or a random population. In some embodiments, said method comprises the additional steps of determining the phenotype in said trait positive and said control populations prior to step c).

Interaction Analysis

The biallelic markers of the present invention may also be used to identify patterns of biallelic markers associated with detectable traits resulting from polygenic interactions. The analysis of genetic interaction between alleles at unlinked loci requires individual genotyping using the techniques described herein. The analysis of allelic interaction among a selected set of biallelic markers with appropriate level of statistical significance can be considered as a haplotype analysis. Interaction analysis consists in stratifying the case-control populations with respect to a given haplotype for the first loci and performing a haplotype analysis with the second loci with each subpopulation.

Statistical methods used in association studies are further described below.

**Testing For Linkage In The Presence Of Association**

The biallelic markers of the present invention may further be used in TDT (transmission/disequilibrium test). TDT tests for both linkage and association and is not affected by population stratification. TDT requires data for affected individuals and their parents or data from unaffected sibs instead of from parents (see Spielmann S. et al., 1993; Schaid D.J. et al., 1996, Spielmann S. and Ewens W.J., 1998). Such combined tests generally reduce the false – positive errors produced by separate analyses.

**Association OF Biallelic Markers Of *BAP28* With Prostate Cancer**

**Trait Positive And Control Populations**

Two groups of independent individuals were used: the overall trait positive and the control populations included 491 individuals suffering from prostate cancer and 313 individuals without any sign of prostate cancer. A specific protocol for the collection of DNA samples from trait positive and control individuals is described in Example 5. The 491 affected individuals can be subdivided in 197 familial cases and 294 sporadic cases. The sporadic cases comprises 70 sporadic informatives cases. The 491 individuals suffering from prostate cancer can also be subdivided into a population of individuals who developed prostate cancer under 65 years-old and a population of individuals who developed prostate cancer after the age of 65.

In order to have as much certainty as possible on the absence of prostate cancer in control individuals, it is preferred to conduct a PSA dosage analysis on this population. Several commercial assays can be used (WO 96/21042, herein by reference). In one preferred embodiment, a Hybritech assay is used and control individuals must have a level of PSA less than 2.8 ng/ml of serum in order to be selected as such. In a preferred embodiment, the Yang assay is used and trait negative individuals must have a level of PSA of less than 4 ng/ml of serum in order to be included in the population under study. More preferably, the control population is at least 65 year old.

**Association Analysis**

The association analysis showed an association between BAP28-related biallelic markers and prostate cancer, more particularly both familial prostate cancer and sporadic prostate cancer. The results of the association study werefurther details in example 5.

A single point analysis of the association study showed an association between biallelic markers of the *BAP28* gene and prostate cancer, preferably sporadic prostate cancer is associated most strongly with the biallelic markers A28 (5-14/165), A4 (5-382/316), A1 (5-381/133), and A55 (99-7182/49) which present a particular interest (Figures 5 and 6). These association results constitute new elements for studying the genetic susceptibility of individuals to prostate cancer, preferably to sporadic and familial prostate cancer. Further details concerning this association study are provided in Figures 5 and 6 and in the example 5.

Similar association studies can also be carried out with other biallelic markers within the scope of the invention, preferably with biallelic markers in linkage disequilibrium with the markers associated with prostate cancer as described above, including the biallelic markers A1 to A58.

**Haplotype Analysis**

In the context of the present invention, a haplotype can be defined as a combination of biallelic markers found in a given individual and which may be associated more or less significantly, as a result of appropriate statistical analyses, with the expression of a given trait.

The haplotype studies are detailed in Example 5.

Several two-marker haplotypes were significantly associated with familial prostate cancer. One preferred two-marker haplotype including markers A30 (99-1572/440) and A32 (5-171/204), alleles TT respectively, was shown to be significantly associated with prostate cancer, preferably with familial prostate cancer. As shown in Figures 8, 9 and 12 A, this haplotype presents a p-value of $2.5\ 10^{-6}$ for the early onset familial prostate cancer (see Example 5).

Several two-marker haplotypes were significantly associated with sporadic prostate cancer. One preferred two-marker haplotype including markers A16 (5-370/197), and A1 (5-381/133), alleles GA was shown to be significantly associated with sporadic prostate cancer. As shown in Figures 10, 11 and 12 B, this haplotype presents a p-value of $9.4 \times 10^{-8}$ for the informative sporadic prostate cancer (see Example 5).

Several two-marker haplotypes were significantly associated with sporadic prostate cancer. One preferred two-marker haplotype including markers A53 (99-1601/402), and A4 (5-382/316), alleles TG, was shown to be significantly associated with sporadic prostate cancer. As shown in Figures 10, 11 and 12 C, this haplotype presents a p-value of $1 \times 10^{-5}$ for the informative sporadic prostate cancer (see Example 5).

Several three-biallelic marker haplotypes are described in the Example 5.

The permutation tests clearly validated the statistical significance of the association between these haplotypes and the prostate cancer (see Example 5). All these haplotypes can be used in diagnostic of prostate cancer, more particularly either familial prostate cancer or sporadic prostate cancer.

This information is extremely valuable. The knowledge of a potential genetic predisposition to prostate cancer, even if this predisposition is not absolute, might contribute in a very significant manner to treatment efficacy of prostate cancer and to the development of new therapeutic and diagnostic tools.

The invention concerns a haplotype comprising at least one biallelic marker selected from the group consisting of A1 to A58, preferably A54, A58, A57, A56, A55, A1, A4, A5, A7, A11, A12, A16, A19, A21, A25, A27, A28, A29, A35, A33, A34, A32, A31, A30, A50, A51, A42, A53, A43, and A48, more preferably A1, A4, A30, A31, A42, A51, and A53.

**Statistical methods**

In general, any method known in the art to test whether a trait and a genotype show a statistically significant correlation may be used.

**1) Methods In Linkage Analysis**

Statistical methods and computer programs useful for linkage analysis are well-known to those skilled in the art (see Terwilliger J.D. and Ott J., 1994; Ott J., 1991).

**2) Methods To Estimate Haplotype Frequencies In A Population**

As described above, when genotypes are scored, it is often not possible to distinguish heterozygotes so that haplotype frequencies cannot be easily inferred. When the gametic phase is not known, haplotype frequencies can be estimated from the multilocus genotypic data. Any method known to person skilled in the art can be used to estimate haplotype frequencies (see Lange K., 1997; Weir, B.S., 1996) Preferably, maximum-likelihood haplotype frequencies are computed using an Expectation- Maximization (EM) algorithm (see Dempster et al., 1977; Excoffier L. and Slatkin M., 1995). This procedure is an iterative process aiming at obtaining maximum-likelihood estimates of haplotype frequencies from multi-locus genotype data when the gametic phase is unknown. Haplotype estimations are usually performed by applying the EM algorithm using for example the EM-HAPLO program (Hawley M. E. et al., 1994) or the Arlequin program (Schneider et al., 1997). The EM algorithm is a generalized iterative maximum likelihood approach to estimation and is briefly described below.

Please note that in the present section, "Methods To Estimate Haplotype Frequencies In A Population," of this text, phenotypes will refer to multi-locus genotypes with unknown phase. Genotypes will refer to known-phase multi-locus genotypes.

A sample of N unrelated individuals is typed for K markers. The data observed are the unknown-phase K-locus phenotypes that can categorized in F different phenotypes. Suppose that we have H underlying possible haplotypes (in case of K biallelic markers, $H=2^K$).

For phenotype j, suppose that $c_j$ genotypes are possible. We thus have the following equation

$$P_j = \sum_{i=1}^{c_j} pr(genotype_i) = \sum_{i=1}^{c_j} pr(h_k, h_l) \qquad \text{Equation 1}$$

where $Pj$ is the probability of the phenotype $j$, $h_k$ and $h_l$ are the two haplotypes constituent the genotype $i$. Under the Hardy-Weinberg equilibrium, $pr(h_k,h_l)$ becomes:

$$pr(h_k, h_l) = pr(h_k)^2 \text{ if } h_k = h_l,\ pr(h_k, h_l) = 2\,pr(h_k).pr(h_l) \text{ if } h_k \neq h_l. \qquad \text{Equation 2}$$

The successive steps of the E-M algorithm can be described as follows:

Starting with initial values of the of haplotypes frequencies, noted $p_1^{(0)}, p_2^{(0)} \ldots\ldots p_H^{(0)}$, these initial values serve to estimate the genotype frequencies (Expectation step) and then estimate

another set of haplotype frequencies (Maximization step), noted $p_1^{(1)}, p_2^{(1)}, \ldots p_H^{(1)}$, these two steps are iterated until changes in the sets of haplotypes frequency are very small.

A stop criterion can be that the maximum difference between haplotype frequencies between two iterations is less than $10^{-3}$. These values can be adjusted according to the desired precision of estimations.

At a given iteration s, the Expectation step consists in calculating the genotypes frequencies by the following equation:

$$pr(genotype_i)^{(s)} = pr(phenotype_j).pr(genotype_i | phenotype_j)^{(s)}$$

$$= \frac{n_j}{N} \cdot \frac{pr(h_k, h_l)^{(s)}}{P_j^{(s)}} \qquad \text{Equation 3}$$

where genotype $i$ occurs in phenotype $j$, and where $h_k$ and $h_l$ constitute genotype $i$. Each probability is derived according to eq. 1, and eq. 2 described above.

Then the Maximization step simply estimates another set of haplotype frequencies given the genotypes frequencies. This approach is also known as the gene-counting method (Smith, 1957).

$$p_t^{(s+1)} = \frac{1}{2} \sum_{j=1}^{F} \sum_{i=1}^{c_j} \delta_{it} \cdot pr(genotype_i)^{(s)} \qquad \text{Equation 4}$$

Where $\delta_{it}$ is an indicator variable which count the number of time haplotype $t$ in genotype $i$. It takes the values of 0, 1 or 2.

To ensure that the estimation finally obtained is the maximum-likelihood estimation several values of departures are required. The estimations obtained are compared and if they are different the estimations leading to the best likelihood are kept.

**3) Methods To Calculate Linkage Disequilibrium Between Markers**

A number of methods can be used to calculate linkage disequilibrium between any two genetic positions, in practice linkage disequilibrium is measured by applying a statistical association test to haplotype data taken from a population.

Linkage disequilibrium between any pair of biallelic markers comprising at least one of the biallelic markers of the present invention ($M_i$, $M_j$) having alleles ($a_i$/$b_i$) at marker $M_i$ and alleles ($a_j$/$b_j$) at marker $M_j$ can be calculated for every allele combination ($a_i$,$a_j$, $a_i$,$b_j$, $b_i$,$a_j$ and $b_i$,$b_j$), according to the Piazza formula:

$\Delta_{ij,kl} = \sqrt{\theta 4} - \sqrt{(\theta 4 + \theta 3)(\theta 4 + \theta 2)}$, where:

$\theta 4$ = - - = frequency of genotypes not having allele $a_i$ at $M_i$ and not having allele $a_j$ at $M_j$

$\theta 3$ = - + = frequency of genotypes not having allele $a_i$ at $M_i$ and having allele $a_j$ at $M_j$

$\theta 2$ = + - = frequency of genotypes having allele $a_i$ at $M_i$ and not having allele $a_j$ at $M_j$

Linkage disequilibrium (LD) between pairs of biallelic markers ($M_i$, $M_j$) can also be calculated for every allele combination (ai,aj, ai,bj, $b_i$,$a_j$ and $b_i$,$b_j$), according to the maximum-likelihood estimate (MLE) for delta (the composite genotypic disequilibrium coefficient), as described by Weir (Weir B. S., 1996). The MLE for the composite linkage disequilibrium is:

$D_{aiaj} = (2n_1 + n_2 + n_3 + n_4/2)/N - 2(pr(a_i).\ pr(a_j))$

Where $n_1 = \Sigma$ phenotype ($a_i/a_i$, $a_j/a_j$), $n_2 = \Sigma$ phenotype ($a_i/a_i$, $a_j/b_j$), $n_3 = \Sigma$ phenotype ($a_i/b_i$, $a_j/a_j$), n4= $\Sigma$ phenotype ($a_i/b_i$, $a_j/b_j$) and N is the number of individuals in the sample.

This formula allows linkage disequilibrium between alleles to be estimated when only genotype, and not haplotype, data are available.

Another means of calculating the linkage disequilibrium between markers is as follows. For a couple of biallelic markers, $M_i$ ($a_i/b_i$) and $M_j$ ($a_j/b_j$), fitting the Hardy-Weinberg equilibrium, one can estimate the four possible haplotype frequencies in a given population according to the approach described above.

The estimation of gametic disequilibrium between *ai* and *aj* is simply:

$$D_{aiaj} = pr(haplotype(a_i, a_j)) - pr(a_i).pr(a_j).$$

Where $pr(a_i)$ is the probability of allele $a_i$ and $pr(a_j)$ is the probability of allele $a_j$ and where $pr(haplotype\ (a_i, a_j))$ is estimated as in Equation 3 above.

For a couple of biallelic marker only one measure of disequilibrium is necessary to describe the association between $M_i$ and $M_j$.

Then a normalized value of the above is calculated as follows:

$D'_{aiaj} = D_{aiaj} / \max(-pr(a_i).\ pr(a_j),\ -pr(b_i).\ pr(b_j))$ with $D_{aiaj}<0$

$D'_{aiaj} = D_{aiaj} / \max(pr(b_i).\ pr(a_j),\ pr(a_i).\ pr(b_j))$ with $D_{aiaj}>0$

The skilled person will readily appreciate that other linkage disequilibrium calculation methods can be used.

Linkage disequilibrium among a set of biallelic markers having an adequate heterozygosity rate can be determined by genotyping between 50 and 1000 unrelated individuals, preferably between 75 and 200, more preferably around 100.

**4) Testing For Association**

Methods for determining the statistical significance of a correlation between a phenotype and a genotype, in this case an allele at a biallelic marker or a haplotype made up of such alleles, may be determined by any statistical test known in the art and with any accepted threshold of statistical significance being required. The application of particular methods and thresholds of significance are well with in the skill of the ordinary practitioner of the art.

Testing for association is performed by determining the frequency of a biallelic marker allele in case and control populations and comparing these frequencies with a statistical test to determine if their is a statistically significant difference in frequency which would indicate a correlation between the trait and the biallelic marker allele under study. Similarly, a haplotype analysis is performed by estimating the frequencies of all possible haplotypes for a given set of biallelic markers in case and control populations, and comparing these frequencies with a statistical test to determine if their is a statistically significant correlation between the haplotype and the phenotype (trait) under study. Any statistical tool useful to test for a statistically significant association between a genotype and a phenotype may be used. Preferably the statistical test employed is a chi-square test with one degree of freedom. A P-value is calculated (the P-value is the probability that a statistic as large or larger than the observed one would occur by chance).

Statistical Significance

In preferred embodiments, significance for diagnosis purposes, either as a positive basis for further diagnostic tests or as a preliminary starting point for early preventive therapy, the p value related to a biallelic marker association is preferably about $1 \times 10^{-2}$ or less, more preferably about $1 \times 10^{-4}$ or less, for a single biallelic marker analysis and about $1 \times 10^{-3}$ or less, still more preferably $1 \times 10^{-6}$ or less and most preferably of about $1 \times 10^{-8}$ or less, for a haplotype analysis involving two or more markers. These values are believed to be applicable to any association studies involving single or multiple marker combinations.

The skilled person can use the range of values set forth above as a starting point in order to carry out association studies with biallelic markers of the present invention. In doing so, significant associations between the biallelic markers of the present invention and a trait can be revealed and used for diagnosis and drug screening purposes.

Phenotypic Permutation

In order to confirm the statistical significance of the first stage haplotype analysis described above, it might be suitable to perform further analyses in which genotyping data from case-control individuals are pooled and randomized with respect to the trait phenotype. Each individual genotyping data is randomly allocated to two groups, which contain the same number of individuals as the case-control populations used to compile the data obtained in the first stage. A second stage haplotype analysis is preferably run on these artificial groups, preferably for the markers included in the haplotype of the first stage analysis showing the highest relative risk coefficient. This experiment is reiterated preferably at least between 100 and 10000 times. The repeated iterations allow the determination of the probability to obtain the tested haplotype by chance.

Assessment Of Statistical Association

To address the problem of false positives similar analysis may be performed with the same case-control populations in random genomic regions. Results in random regions and the candidate

region are compared as described in a co-pending US Provisional Patent Application entitled "Methods, Software And Apparati For Identifying Genomic Regions Harboring A Gene Associated With A Detectable Trait," U.S. Serial Number 60/107,986, filed November 10, 1998, the contents of which are incorporated herein by reference.

**5) Evaluation Of Risk Factors**

The association between a risk factor (in genetic epidemiology the risk factor is the presence or the absence of a certain allele or haplotype at marker loci) and a disease is measured by the odds ratio (OR) and by the relative risk (RR). If $P(R^+)$ is the probability of developing the disease for individuals with R and $P(R^-)$ is the probability for individuals without the risk factor, then the relative risk is simply the ratio of the two probabilities, that is:

$RR = P(R^+)/P(R^-)$

In case-control studies, direct measures of the relative risk cannot be obtained because of the sampling design. However, the odds ratio allows a good approximation of the relative risk for low-incidence diseases and can be calculated:

$$OR = \left[ \begin{array}{c} F^+ \\ 1 - F^+ \end{array} \right] \Big/ \left[ \begin{array}{c} F^- \\ (1 - F^-) \end{array} \right]$$

$OR = (F^+/(1-F^+))/(F^-/(1-F^-))$

$F^+$ is the frequency of the exposure to the risk factor in cases and $F^-$ is the frequency of the exposure to the risk factor in controls. $F^+$ and $F^-$ are calculated using the allelic or haplotype frequencies of the study and further depend on the underlying genetic model (dominant, recessive, additive...).

One can further estimate the attributable risk (AR) which describes the proportion of individuals in a population exhibiting a trait due to a given risk factor. This measure is important in quantifying the role of a specific factor in disease etiology and in terms of the public health impact of a risk factor. The public health relevance of this measure lies in estimating the proportion of cases of disease in the population that could be prevented if the exposure of interest were absent. AR is determined as follows:

$AR = P_E (RR-1) / (P_E (RR-1)+1)$

AR is the risk attributable to a biallelic marker allele or a biallelic marker haplotype. $P_E$ is the frequency of exposure to an allele or a haplotype within the population at large; and RR is the relative risk which, is approximated with the odds ratio when the trait under study has a relatively low incidence in the general population.

**Identification Of Biallelic Markers In Linkage Disequilibrium With The Biallelic Markers of the Invention**

Once a first biallelic marker has been identified in a genomic region of interest, the practitioner of ordinary skill in the art, using the teachings of the present invention, can easily

identify additional biallelic markers in linkage disequilibrium with this first marker. As mentioned before any marker in linkage disequilibrium with a first marker associated with a trait will be associated with the trait. Therefore, once an association has been demonstrated between a given biallelic marker and a trait, the discovery of additional biallelic markers associated with this trait is of great interest in order to increase the density of biallelic markers in this particular region. The causal gene or mutation will be found in the vicinity of the marker or set of markers showing the highest correlation with the trait.

Identification of additional markers in linkage disequilibrium with a given marker involves: (a) amplifying a genomic fragment comprising a first biallelic marker from a plurality of individuals; (b) identifying of second biallelic markers in the genomic region harboring said first biallelic marker; (c) conducting a linkage disequilibrium analysis between said first biallelic marker and second biallelic markers; and (d) selecting said second biallelic markers as being in linkage disequilibrium with said first marker. Subcombinations comprising steps (b) and (c) are also contemplated.

Methods to identify biallelic markers and to conduct linkage disequilibrium analysis are described herein and can be carried out by the skilled person without undue experimentation. The present invention then also concerns biallelic markers which are in linkage disequilibrium with the specific biallelic markers A1 to A58, preferably one of the biallelic markers A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one of, the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53, and which are expected to present similar characteristics in terms of their respective association with a given trait. In a preferred embodiment, the invention concerns biallelic markers which are in linkage disequilibrium with the specific biallelic markers

**Identification Of Functional Mutations**

Mutations in the *BAP28* gene which are responsible for a detectable phenotype or trait may be identified by comparing the sequences of the *BAP28* gene from trait positive and control individuals. Once a positive association is confirmed with a biallelic marker of the present invention, the identified locus can be scanned for mutations. In a preferred embodiment, functional regions such as exons and splice sites, promoters and other regulatory regions of the *BAP28* gene are scanned for mutations. In a preferred embodiment the sequence of the *BAP28* gene is compared in trait positive and control individuals. Preferably, trait positive individuals carry the haplotype shown to be associated with the trait and trait negative individuals do not carry the haplotype or allele associated with the trait. The detectable trait or phenotype may comprise a variety of manifestations of altered *BAP28* function.

The mutation detection procedure is essentially similar to that used for biallelic marker identification. The method used to detect such mutations generally comprises the following steps:

- amplification of a region of the *BAP28* gene comprising a biallelic marker or a group of biallelic markers associated with the trait from DNA samples of trait positive patients and trait-negative controls;

- sequencing of the amplified region;

- comparison of DNA sequences from trait positive and control individuals;

- determination of mutations specific to trait-positive patients.

In one embodiment, said biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58. In a more preferred embodiment, said biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53. It is preferred that candidate polymorphisms be then verified by screening a larger population of cases and controls by means of any genotyping procedure such as those described herein, preferably using a microsequencing technique in an individual test format. Polymorphisms are considered as candidate mutations when present in cases and controls at frequencies compatible with the expected association results. Polymorphisms are considered as candidate "trait-causing" mutations when they exhibit a statistically significant correlation with the detectable phenotype.

**Biallelic Markers Of The Invention In Methods Of Genetic Diagnostics**

The biallelic markers of the present invention can also be used to develop diagnostics tests capable of identifying individuals who express a detectable trait as the result of a specific genotype or individuals whose genotype places them at risk of developing a detectable trait at a subsequent time. The trait analyzed using the present diagnostics may be any detectable trait, including susceptibility to prostate cancer, the level of aggressiveness of prostate cancer tumors, an early onset of prostate cancer, a beneficial response to or side effects related to treatment against prostate cancer. Such a diganosis can be useful in the staging, monitoring, prognosis and/or prophylactic or curative therapy of prostate cancer.

The diagnostic techniques of the present invention may employ a variety of methodologies to determine whether a test subject has a biallelic marker pattern associated with an increased risk of developing a detectable trait or whether the individual suffers from a detectable trait as a result of a particular mutation, including methods which enable the analysis of individual chromosomes for haplotyping, such as family studies, single sperm DNA analysis or somatic hybrids.

The present invention provides diagnostic methods to determine whether an individual is at risk of developing a disease or suffers from a disease resulting from a mutation or a polymorphism in the *BAP28* gene. The present invention also provides methods to determine whether an individual has a susceptibility to prostate cancer.

These methods involve obtaining a nucleic acid sample from the individual and, determining, whether the nucleic acid sample contains at least one allele or at least one biallelic

marker haplotype, indicative of a risk of developing the trait or indicative that the individual expresses the trait as a result of possessing a particular *BAP28* polymorphism or mutation (trait-causing allele).

Preferably, in such diagnostic methods, a nucleic acid sample is obtained from the individual and this sample is genotyped using methods described above in "Methods Of Genotyping DNA Samples For Biallelic markers. The diagnostics may be based on a single biallelic marker or a on group of biallelic markers.

In each of these methods, a nucleic acid sample is obtained from the test subject and the biallelic marker pattern of one or more of the biallelic markers A1 to A58, preferably one or more of the biallelic markers A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one or more of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53, is determined.

In one embodiment, a PCR amplification is conducted on the nucleic acid sample to amplify regions in which polymorphisms associated with a detectable phenotype have been identified. The amplification products are sequenced to determine whether the individual possesses one or more *BAP28* polymorphisms associated with a detectable phenotype. The primers used to generate amplification products may comprise the primers listed in Table 1. Alternatively, the nucleic acid sample is subjected to microsequencing reactions as described above to determine whether the individual possesses one or more *BAP28* polymorphisms associated with a detectable phenotype resulting from a mutation or a polymorphism in the *BAP28* gene. The primers used in the microsequencing reactions may include the primers listed in Table 4. In another embodiment, the nucleic acid sample is contacted with one or more allele specific oligonucleotide probes which, specifically hybridize to one or more *BAP28* alleles associated with a detectable phenotype. The probes used in the hybridization assay may include the probes listed in Table 3. In another embodiment, the nucleic acid sample is contacted with a second *BAP28* oligonucleotide capable of producing an amplification product when used with the allele specific oligonucleotide in an amplification reaction. The presence of an amplification product in the amplification reaction indicates that the individual possesses one or more *BAP28* alleles associated with a detectable phenotype.

In a preferred embodiment the identity of the nucleotide present at, at least one, biallelic marker selected from the group consisting of A1 to A58 and the complements thereof, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof, is determined and the detectable trait is prostate cancer, more preferably sporadic prostate cancer. Diagnostic kits comprise any of the polynucleotides of the present invention.

These diagnostic methods are extremely valuable as they can, in certain circumstances, be used to initiate preventive treatments or to allow an individual carrying a significant haplotype to foresee warning signs such as minor symptoms.

Diagnostics, which analyze and predict response to a drug or side effects to a drug, may be used to determine whether an individual should be treated with a particular drug. For example, if the diagnostic indicates a likelihood that an individual will respond positively to treatment with a particular drug, the drug may be administered to the individual. Conversely, if the diagnostic indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects.

Clinical drug trials represent another application for the markers of the present invention. One or more markers indicative of response to an agent acting against prostate cancer or to side effects to an agent acting against prostate cancer may be identified using the methods described above. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and without risking undesirable safety problems.

**Treatment Of Prostate Cancer**

As the metastasis of prostate cancer can be fatal, it is important to detect prostate cancer susceptibility of individuals. Consequently, the invention also concerns a method for the treatment of prostate cancer comprising the following steps:

- selecting an individual whose DNA comprises alleles of a biallelic marker or of a group of biallelic markers, preferably *BAP28*-related markers, associated with prostate cancer;
- following up said individual for the appearance (and optionally the development) of tumors in prostate; and
- administering an effective amount of a medicament acting against prostate cancer to said individual at an appropriate stage of the prostate cancer.

In one embodiment, said biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58 and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof.

The prophylactic administration of a treatment serves to prevent, attenuate or inhibit the growth of cancer cells.

Another embodiment of the present invention consists of a method for the treatment of prostate cancer comprising the following steps:

- selecting an individual whose DNA comprises alleles of a biallelic marker or of a group of biallelic markers, preferably *BAP28*-related markers, associated with prostate cancer;
- administering to said individual, preferably as a preventive treatment of prostate cancer, an effective amount of a medicament acting against prostate cancer such as 4HPR.

In one embodiment, said biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58 and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof.

In a further embodiment, the present invention concerns a method for the treatment of prostate cancer comprising the following steps:

- selecting an individual whose DNA comprises alleles of a biallelic marker or of a group of biallelic markers, preferably *BAP28*-related markers, associated with a susceptibility prostate cancer;
- administering to said individual, as a preventive treatment of prostate cancer, an effective amount of a medicament acting against prostate cancer such as 4HPR;
- following up said individual for the appearance and the development of tumors in prostate; and optionally
- administering an effective amount of a medicament acting against prostate cancer to said individual at the appropriate stage of the prostate cancer.

In one embodiment, said biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58 and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof.

To enlighten the choice of the appropriate beginning of the treatment of prostate cancer, the present invention also concerns a method for the treatment of prostate cancer comprising the following steps:

- selecting an individual suffering from a prostate cancer whose DNA comprises alleles of a biallelic marker or of a group of biallelic markers, preferably *BAP28*-related markers, associated with the aggressiveness of prostate cancer tumors; and
- administering an effective amount of a medicament acting against prostate cancer to said individual.

In one embodiment, said biallelic marker is selected from the group consisting of A1 to A58, and the complements thereof. In a preferred embodiment, said biallelic marker is selected from

the group consisting of A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58 and the complements thereof. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof. In particular embodiments, the individual is selected by genotyping one or more biallelic markers of the present invention.

**Recombinant Vectors**

The term "vector" is used herein to designate either a circular or a linear DNA or RNA molecule, which is either double-stranded or single-stranded, and which comprise at least one polynucleotide of interest that is sought to be transferred in a cell host or in a unicellular or multicellular host organism.

The present invention encompasses a family of recombinant vectors that comprise a regulatory polynucleotide derived from the *BAP28* genomic sequence, and/or a coding polynucleotide from either the *BAP28* genomic sequence or the cDNA sequence.

Generally, a recombinant vector of the invention may comprise any of the polynucleotides described herein, including regulatory sequences, coding sequences and polynucleotide constructs, as well as any *BAP28* primer or probe as defined above. More particularly, the recombinant vectors of the present invention can comprise any of the polynucleotides described in the "Genomic Sequences Of The *BAP28* Gene" section, the " *BAP28* cDNA Sequences" section, the "Coding Regions" section, the "Polynucleotide constructs" section, and the "Oligonucleotide Probes And Primers" section.

In a first preferred embodiment, a recombinant vector of the invention is used to amplify the inserted polynucleotide derived from a *BAP28* genomic sequence of SEQ ID No 1 or a *BAP28* cDNA, for example the cDNA of SEQ ID No 2, 3 or 4 in a suitable cell host, this polynucleotide being amplified at every time that the recombinant vector replicates.

A second preferred embodiment of the recombinant vectors according to the invention consists of expression vectors comprising either a regulatory polynucleotide or a coding nucleic acid of the invention, or both. Within certain embodiments, expression vectors are employed to express the BAP28 polypeptide which can be then purified and, for example be used in ligand screening assays or as an immunogen in order to raise specific antibodies directed against the BAP28 protein. In other embodiments, the expression vectors are used for constructing transgenic animals and also for gene therapy. Expression requires that appropriate signals are provided in the vectors, said signals including various regulatory elements, such as enhancers/promoters from both viral and mammalian sources that drive expression of the genes of interest in host cells. Dominant drug selection markers for establishing permanent, stable cell clones expressing the products are generally included in the expression vectors of the invention, as they are elements that link expression of the drug selection markers to expression of the polypeptide.

In a further embodiment, the invention concerns a vector comprising a polynucleotide sequence seleted from the group consisting of SEQ ID Nos 4, and 9-13, a complementary sequence thereto or a fragment thereof.

More particularly, the present invention relates to expression vectors which include nucleic acids encoding a BAP28 protein, preferably the BAP28 protein of the amino acid sequence of SEQ ID No 5 or variants or fragments thereof.

The invention also pertains to a recombinant expression vector useful for the expression of the *BAP28* coding sequence, wherein said vector comprises a nucleic acid of SEQ ID No 2 or 3.

Recombinant vectors comprising a nucleic acid containing a *BAP28*-related biallelic marker is also part of the invention. In a preferred embodiment, said biallelic marker is selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof.

Some of the elements which can be found in the vectors of the present invention are described in further detail in the following sections.

The present invention also encompasses primary, secondary, and immortalized homologously recombinant host cells of vertebrate origin, preferably mammalian origin and particularly human origin, that have been engineered to: a) insert exogenous (heterologous) polynucleotides into the endogenous chromosomal DNA of a targeted gene, b) delete endogenous chromosomal DNA, and/or c) replace endogenous chromosomal DNA with exogenous polynucleotides. Insertions, deletions, and/or replacements of polynucleotide sequences may be to the coding sequences of the targeted gene and/or to regulatory regions, such as promoter and enhancer sequences, operably associated with the targeted gene.

The present invention further relates to a method of making a homologously recombinant host cell in vitro or in vivo, wherein the expression of a targeted gene not normally expressed in the cell is altered. Preferably the alteration causes expression of the targeted gene under normal growth conditions or under conditions suitable for producing the polypeptide encoded by the targeted gene. The method comprises the steps of: (a) transfecting the cell in vitro or in vivo with a polynucleotide construct, the a polynucleotide construct comprising: (i) a targeting sequence; (ii) a regulatory sequence and/or a coding sequence; and (iii) an unpaired splice donor site, if necessary, thereby producing a transfected cell; and (b) maintaining the transfected cell in vitro or in vivo under conditions appropriate for homologous recombination.

The present invention further relates to a method of altering the expression of a targeted gene in a cell in vitro or in vivo wherein the gene is not normally expressed in the cell, comprising the steps of: (a) transfecting the cell in vitro or in vivo with a polynucleotide construct, the a polynucleotide construct comprising: (i) a targeting sequence; (ii) a regulatory sequence and/or a coding sequence; and (iii) an unpaired splice donor site, if necessary, thereby producing a

transfected cell; and (b) maintaining the transfected cell in vitro or in vivo under conditions appropriate for homologous recombination, thereby producing a homologously recombinant cell; and (c) maintaining the homologously recombinant cell in vitro or in vivo under conditions appropriate for expression of the gene.

The present invention further relates to a method of making a polypeptide of the present invention by altering the expression of a targeted endogenous gene in a cell in vitro or in vivo wherein the gene is not normally expressed in the cell, comprising the steps of: a) transfecting the cell in vitro with a a polynucleotide construct, the a polynucleotide construct comprising: (i) a targeting sequence; (ii) a regulatory sequence and/or a coding sequence; and (iii) an unpaired splice donor site, if necessary, thereby producing a transfected cell; (b) maintaining the transfected cell in vitro or in vivo under conditions appropriate for homologous recombination, thereby producing a homologously recombinant cell; and c) maintaining the homologously recombinant cell in vitro or in vivo under conditions appropriate for expression of the gene thereby making the polypeptide.

The present invention further relates to a polynucleotide construct which alters the expression of a targeted gene in a cell type in which the gene is not normally expressed. This occurs when the a polynucleotide construct is inserted into the chromosomal DNA of the target cell, wherein the a polynucleotide construct comprises: a) a targeting sequence; b) a regulatory sequence and/or coding sequence; and c) an unpaired splice-donor site, if necessary. Further included are a polynucleotide constructs, as described above, wherein the construct further comprises a polynucleotide which encodes a polypeptide and is in-frame with the targeted endogenous gene after homologous recombination with chromosomal DNA.

The compositions may be produced, and methods performed, by techniques known in the art, such as those described in U.S. Patent Nos: 6,054,288; 6,048,729; 6,048,724; 6,048,524; 5,994,127; 5,968,502; 5,965,125; 5,869,239; 5,817,789; 5,783,385; 5,733,761; 5,641,670; 5,580,734 ; International Publication Nos:WO96/29411, WO 94/12650; and scientific articles including 1994; Koller et al., Proc. Natl. Acad. Sci. USA 86:8932-8935 (1989) (the disclosures of each of which are incorporated by reference in their entireties).

**1. General features of the expression vectors of the invention**

A recombinant vector according to the invention comprises, but is not limited to, a YAC (Yeast Artificial Chromosome), a BAC (Bacterial Artificial Chromosome), a phage, a phagemid, a cosmid, a plasmid or even a linear DNA molecule which may comprise a chromosomal, non-chromosomal, semi-synthetic and synthetic DNA. Such a recombinant vector can comprise a transcriptional unit comprising an assembly of:

(1) a genetic element or elements having a regulatory role in gene expression, for example promoters or enhancers. Enhancers are cis-acting elements of DNA, usually from about 10 to 300 bp in length that act on the promoter to increase the transcription.

(2) a structural or coding sequence which is transcribed into mRNA and eventually translated into a polypeptide, said structural or coding sequence being operably linked to the regulatory elements described in (1); and

(3) appropriate transcription initiation and termination sequences. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling extracellular secretion of translated protein by a host cell. Alternatively, when a recombinant protein is expressed without a leader or transport sequence, it may include a N-terminal residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

Generally, recombinant expression vectors will include origins of replication, selectable markers permitting transformation of the host cell, and a promoter derived from a highly expressed gene to direct transcription of a downstream structural sequence. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably a leader sequence capable of directing secretion of the translated protein into the periplasmic space or the extracellular medium. In a specific embodiment wherein the vector is adapted for transfecting and expressing desired sequences in mammalian host cells, preferred vectors will comprise an origin of replication in the desired host, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5`-flanking non-transcribed sequences. DNA sequences derived from the SV40 viral genome, for example SV40 origin, early promoter, enhancer, splice and polyadenylation sites may be used to provide the required non-transcribed genetic elements.

The *in vivo* expression of a BAP28 polypeptide of SEQ ID No 5 or fragments or variants thereof may be useful in order to correct a genetic defect related to the expression of the native gene in a host organism or to the production of a biologically inactive BAP28 protein.

Consequently, the present invention also deals with recombinant expression vectors mainly designed for the *in vivo* production of the BAP28 polypeptide of SEQ ID No 5 or fragments or variants thereof by the introduction of the appropriate genetic material in the organism of the patient to be treated. This genetic material may be introduced *in vitro* in a cell that has been previously extracted from the organism, the modified cell being subsequently reintroduced in the said organism, directly *in vivo* into the appropriate tissue.

**2. Regulatory Elements**

Promoters

The suitable promoter regions used in the expression vectors according to the present invention are chosen taking into account the cell host in which the heterologous gene has to be expressed. The particular promoter employed to control the expression of a nucleic acid sequence of interest is not believed to be important, so long as it is capable of directing the expression of the

nucleic acid in the targeted cell. Thus, where a human cell is targeted, it is preferable to position the nucleic acid coding region adjacent to and under the control of a promoter that is capable of being expressed in a human cell, such as, for example, a human or a viral promoter.

A suitable promoter may be heterologous with respect to the nucleic acid for which it controls the expression or alternatively can be endogenous to the native polynucleotide containing the coding sequence to be expressed. Additionally, the promoter is generally heterologous with respect to the recombinant vector sequences within which the construct promoter/coding sequence has been inserted.

Promoter regions can be selected from any desired gene using, for example, CAT (chloramphenicol transferase) vectors and more preferably pKK232-8 and pCM7 vectors.

Preferred bacterial promoters are the LacI, LacZ, the T3 or T7 bacteriophage RNA polymerase promoters, the gpt, lambda PR, PL and trp promoters (EP 0036776), the polyhedrin promoter, or the p10 protein promoter from baculovirus (Kit Novagen) (Smith et al., 1983; O'Reilly et al., 1992), the lambda PR promoter or also the trc promoter.

Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of a convenient vector and promoter is well within the level of ordinary skill in the art.

The choice of a promoter is well within the ability of a person skilled in the field of genetic egineering. For example, one may refer to the book of Sambrook et al.(1989) or also to the procedures described by Fuller et al.(1996).

Other regulatory elements

Where a cDNA insert is employed, one will typically desire to include a polyadenylation signal to effect proper polyadenylation of the gene transcript. The nature of the polyadenylation signal is not believed to be crucial to the successful practice of the invention, and any such sequence may be employed such as human growth hormone and SV40 polyadenylation signals. Also contemplated as an element of the expression cassette is a terminator. These elements can serve to enhance message levels and to minimize read through from the cassette into other sequences.

**3. Selectable Markers**

Such markers would confer an identifiable change to the cell permitting easy identification of cells containing the expression construct. The selectable marker genes for selection of transformed host cells are preferably dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, TRP1 for *S. cerevisiae* or tetracycline, rifampicin or ampicillin resistance in *E. coli*, or levan saccharase for mycobacteria, this latter marker being a negative selection marker.

**4. Preferred Vectors.**

Bacterial vectors

As a representative but non-limiting example, useful expression vectors for bacterial use can comprise a selectable marker and a bacterial origin of replication derived from commercially available plasmids comprising genetic elements of pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia, Uppsala, Sweden), and GEM1 (Promega Biotec, Madison, WI, USA).

Large numbers of other suitable vectors are known to those of skill in the art, and commercially available, such as the following bacterial vectors: pQE70, pQE60, pQE-9 (Qiagen), pbs, pD10, phagescript, psiX174, pbluescript SK, pbsks, pNH8A, pNH16A, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene); pSVK3, pBPV, pMSG, pSVL (Pharmacia); pQE-30 (QIAexpress).

Bacteriophage vectors

The P1 bacteriophage vector may contain large inserts ranging from about 80 to about 100 kb.

The construction of P1 bacteriophage vectors such as p158 or p158/neo8 are notably described by Sternberg (1992, 1994). Recombinant P1 clones comprising *BAP28* nucleotide sequences may be designed for inserting large polynucleotides of more than 40 kb (Linton et al., 1993). To generate P1 DNA for transgenic experiments, a preferred protocol is the protocol described by McCormick et al.(1994). Briefly, *E. coli* (preferably strain NS3529) harboring the P1 plasmid are grown overnight in a suitable broth medium containing 25 µg/ml of kanamycin. The P1 DNA is prepared from the *E. coli* by alkaline lysis using the Qiagen Plasmid Maxi kit (Qiagen, Chatsworth, CA, USA), according to the manufacturer's instructions. The P1 DNA is purified from the bacterial lysate on two Qiagen-tip 500 columns, using the washing and elution buffers contained in the kit. A phenol/chloroform extraction is then performed before precipitating the DNA with 70% ethanol. After solubilizing the DNA in TE (10 mM Tris-HCl, pH 7.4, 1 mM EDTA), the concentration of the DNA is assessed by spectrophotometry.

When the goal is to express a P1 clone comprising *BAP28* nucleotide sequences in a transgenic animal, typically in transgenic mice, it is desirable to remove vector sequences from the P1 DNA fragment, for example by cleaving the P1 DNA at rare-cutting sites within the P1 polylinker (*Sfi*I, *Not*I or *Sal*I). The P1 insert is then purified from vector sequences on a pulsed-field agarose gel, using methods similar using methods similar to those originally reported for the isolation of DNA from YACs (Schedl et al., 1993a; Peterson et al., 1993). At this stage, the resulting purified insert DNA can be concentrated, if necessary, on a Millipore Ultrafree-MC Filter Unit (Millipore, Bedford, MA, USA – 30,000 molecular weight limit) and then dialyzed against microinjection buffer (10 mM Tris-HCl, pH 7.4; 250 µM EDTA) containing 100 mM NaCl, 30 µM

spermine, 70 µM spermidine on a microdyalisis membrane (type VS, 0.025 µM from Millipore). The intactness of the purified P1 DNA insert is assessed by electrophoresis on 1% agarose (Sea Kem GTG; FMC Bio-products) pulse-field gel and staining with ethidium bromide.

Baculovirus vectors

A suitable vector for the expression of the BAP28 polypeptide of SEQ ID No 5 or fragments or variants thereof is a baculovirus vector that can be propagated in insect cells and in insect cell lines. A specific suitable host vector system is the pVL1392/1393 baculovirus transfer vector (Pharmingen) that is used to transfect the SF9 cell line (ATCC N°CRL 1711) which is derived from *Spodoptera frugiperda*.

Other suitable vectors for the expression of the BAP28 polypeptide of SEQ ID No 5 or fragments or variants thereof in a baculovirus expression system include those described by Chai et al.(1993), Vlasak et al.(1983) and Lenhard et al.(1996).

Viral vectors

In one specific embodiment, the vector is derived from an adenovirus. Preferred adenovirus vectors according to the invention are those described by Feldman and Steg (1996) or Ohno et al.(1994). Another preferred recombinant adenovirus according to this specific embodiment of the present invention is the human adenovirus type 2 or 5 (Ad 2 or Ad 5) or an adenovirus of animal origin ( French patent application N° FR-93.05954).

Retrovirus vectors and adeno-associated virus vectors are generally understood to be the recombinant gene delivery systems of choice for the transfer of exogenous polynucleotides *in vivo* , particularly to mammals, including humans. These vectors provide efficient delivery of genes into cells, and the transferred nucleic acids are stably integrated into the chromosomal DNA of the host.

Particularly preferred retroviruses for the preparation or construction of retroviral *in vitro* or *in vitro* gene delivery vehicles of the present invention include retroviruses selected from the group consisting of Mink-Cell Focus Inducing Virus, Murine Sarcoma Virus, Reticuloendotheliosis virus and Rous Sarcoma virus. Particularly preferred Murine Leukemia Viruses include the 4070A and the 1504A viruses, Abelson (ATCC No VR-999), Friend (ATCC No VR-245), Gross (ATCC No VR-590), Rauscher (ATCC No VR-998) and Moloney Murine Leukemia Virus (ATCC No VR-190; PCT Application No WO 94/24298). Particularly preferred Rous Sarcoma Viruses include Bryan high titer (ATCC Nos VR-334, VR-657, VR-726, VR-659 and VR-728). Other preferred retroviral vectors are those described in Roth et al.(1996), PCT Application No WO 93/25234, PCT Application No WO 94/ 06920, Roux et al., 1989, Julan et al., 1992 and Neda et al., 1991.

Yet another viral vector system that is contemplated by the invention consists in the adeno-associated virus (AAV). The adeno-associated virus is a naturally occurring defective virus that requires another virus, such as an adenovirus or a herpes virus, as a helper virus for efficient replication and a productive life cycle (Muzyczka et al., 1992). It is also one of the few viruses that may integrate its DNA into non-dividing cells, and exhibits a high frequency of stable integration

(Flotte et al., 1992; Samulski et al., 1989; McLaughlin et al., 1989). One advantageous feature of AAV derives from its reduced efficacy for transducing primary cells relative to transformed cells.

BAC vectors

The bacterial artificial chromosome (BAC) cloning system (Shizuya et al., 1992) has been developed to stably maintain large fragments of genomic DNA (100-300 kb) in *E. coli*. A preferred BAC vector consists of pBeloBAC11 vector that has been described by Kim et al.(1996). BAC libraries are prepared with this vector using size-selected genomic DNA that has been partially digested using enzymes that permit ligation into either the *Bam* HI or *Hind*III sites in the vector. Flanking these cloning sites are T7 and SP6 RNA polymerase transcription initiation sites that can be used to generate end probes by either RNA transcription or PCR methods. After the construction of a BAC library in *E. coli*, BAC DNA is purified from the host cell as a supercoiled circle. Converting these circular molecules into a linear form precedes both size determination and introduction of the BACs into recipient cells. The cloning site is flanked by two *Not* I sites, permitting cloned segments to be excised from the vector by *Not* I digestion. Alternatively, the DNA insert contained in the pBeloBAC11 vector may be linearized by treatment of the BAC vector with the commercially available enzyme lambda terminase that leads to the cleavage at the unique *cos*N site, but this cleavage method results in a full length BAC clone containing both the insert DNA and the BAC sequences.

**5. Delivery Of The Recombinant Vectors**

In order to effect expression of the polynucleotides and polynucleotide constructs of the invention, these constructs must be delivered into a cell. This delivery may be accomplished *in vitro*, as in laboratory procedures for transforming cell lines, or *in vivo* or *ex vivo*, as in the treatment of certain diseases states.

One mechanism is viral infection where the expression construct is encapsulated in an infectious viral particle.

Several non-viral methods for the transfer of polynucleotides into cultured mammalian cells are also contemplated by the present invention, and include, without being limited to, calcium phosphate precipitation (Graham et al., 1973; Chen et al., 1987), DEAE-dextran (Gopal, 1985), electroporation (Tur-Kaspa et al., 1986; Potter et al., 1984), direct microinjection (Harland et al., 1985), DNA-loaded liposomes (Nicolau et al., 1982; Fraley et al., 1979), and receptor-mediated transfection (Wu and Wu, 1987; 1988). Some of these techniques may be successfully adapted for *in vivo* or *ex vivo* use.

Once the expression polynucleotide has been delivered into the cell, it may be stably integrated into the genome of the recipient cell. This integration may be in the cognate location and orientation via homologous recombination (gene replacement) or it may be integrated in a random, non specific location (gene augmentation). In yet further embodiments, the nucleic acid may be

stably maintained in the cell as a separate, episomal segment of DNA. Such nucleic acid segments or "episomes" encode sequences sufficient to permit maintenance and replication independent of or in synchronization with the host cell cycle.

One specific embodiment for a method for delivering a protein or peptide to the interior of a cell of a vertebrate *in vivo* comprises the step of introducing a preparation comprising a physiologically acceptable carrier and a naked polynucleotide operatively coding for the polypeptide of interest into the interstitial space of a tissue comprising the cell, whereby the naked polynucleotide is taken up into the interior of the cell and has a physiological effect. This is particularly applicable for transfer *in vitro* but it may be applied to *in vivo* as well.

Compositions for use *in vitro* and *in vivo* comprising a "naked" polynucleotide are described in PCT application N° WO 90/11092 (Vical Inc.) and also in PCT application No WO 95/11307 (Institut Pasteur, INSERM, Université d'Ottawa) as well as in the articles of Tacson et al.(1996) and of Huygen et al.(1996).

In still another embodiment of the invention, the transfer of a naked polynucleotide of the invention, including a polynucleotide construct of the invention, into cells may be proceeded with a particle bombardment (biolistic), said particles being DNA-coated microprojectiles accelerated to a high velocity allowing them to pierce cell membranes and enter cells without killing them, such as described by Klein et al.(1987).

In a further embodiment, the polynucleotide of the invention may be entrapped in a liposome (Ghosh and Bacchawat, 1991; Wong et al., 1980; Nicolau et al., 1987)

In a specific embodiment, the invention provides a composition for the *in vivo* production of the BAP28 protein or polypeptide described herein. It comprises a naked polynucleotide operatively coding for this polypeptide, in solution in a physiologically acceptable carrier, and suitable for introduction into a tissue to cause cells of the tissue to express the said protein or polypeptide.

The amount of vector to be injected to the desired host organism varies according to the site of injection. As an indicative dose, it will be injected between 0.1 and 100 µg of the vector in an animal body, preferably a mammal body, for example a mouse body.

In another embodiment of the vector according to the invention, it may be introduced *in vitro* in a host cell, preferably in a host cell previously harvested from the animal to be treated and more preferably a somatic cell such as a muscle cell. In a subsequent step, the cell that has been transformed with the vector coding for the desired BAP28 polypeptide or the desired fragment thereof is reintroduced into the animal body in order to deliver the recombinant protein within the body either locally or systemically.

**Cell Hosts**

Another object of the invention consists of a host cell that has been transformed or transfected with one of the polynucleotides described herein, and in particular a polynucleotide

either comprising a *BAP28* regulatory polynucleotide or the coding sequence of the BAP28 polypeptide of SEQ ID Nos 1, 2, 3 or 4 or a fragment or a variant thereof. Also included are host cells that are transformed (prokaryotic cells) or that are transfected (eukaryotic cells) with a recombinant vector such as one of those described above. More particularly, the cell hosts of the present invention can comprise any of the polynucleotides described in the "Genomic Sequences Of The *BAP28* Gene" section, the "*BAP28* cDNA Sequences" section, the "Coding Regions" section, the "Polynucleotide constructs" section, and the "Oligonucleotide Probes And Primers" section.

A further recombinant cell host according to the invention comprises a polynucleotide containing a biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof.

Preferred host cells used as recipients for the expression vectors of the invention are the following:

a) Prokaryotic host cells: *Escherichia coli* strains (I.E.DH5-α strain), *Bacillus subtilis*, *Salmonella typhimurium*, and strains from species like *Pseudomonas*, *Streptomyces* and *Staphylococcus*.

b) Eukaryotic host cells: HeLa cells (ATCC N°CCL2; N°CCL2.1; N°CCL2.2), Cv 1 cells (ATCC N°CCL70), COS cells (ATCC N°CRL1650; N°CRL1651), Sf-9 cells (ATCC N°CRL1711), C127 cells (ATCC N° CRL-1804), 3T3 (ATCC N° CRL-6361), CHO (ATCC N° CCL-61), human kidney 293. (ATCC N° 45504; N° CRL-1573) and BHK (ECACC N° 84100501; N° 84111301).

c) Other mammalian host cells.

The *BAP28* gene expression in mammalian, and typically human, cells may be rendered defective, or alternatively it may be proceeded with the insertion of a *BAP28* genomic or cDNA sequence with the replacement of the *BAP28* gene counterpart in the genome of an animal cell by a *BAP28* polynucleotide according to the invention. These genetic alterations may be generated by homologous recombination events using specific DNA constructs that have been previously described.

One kind of cell hosts that may be used are mammal zygotes, such as murine zygotes. For example, murine zygotes may undergo microinjection with a purified DNA molecule of interest, for example a purified DNA molecule that has previously been adjusted to a concentration range from 1 ng/ml –for BAC inserts- 3 ng/µl –for P1 bacteriophage inserts- in 10 mM Tris-HCl, pH 7.4, 250 µM EDTA containing 100 mM NaCl, 30 µM spermine, and 70 µM spermidine. When the DNA to be microinjected has a large size, polyamines and high salt concentrations can be used in order to avoid mechanical breakage of this DNA, as described by Schedl et al (1993b).

Anyone of the polynucleotides of the invention, including the DNA constructs described herein, may be introduced in an embryonic stem (ES) cell line, preferably a mouse ES cell line. ES cell lines are derived from pluripotent, uncommitted cells of the inner cell mass of pre-implantation

blastocysts. Preferred ES cell lines are the following: ES-E14TG2a (ATCC n° CRL-1821), ES-D3 (ATCC n° CRL1934 and n° CRL-11632), YS001 (ATCC n° CRL-11776), 36.5 (ATCC n° CRL-11116). To maintain ES cells in an uncommitted state, they are cultured in the presence of growth inhibited feeder cells which provide the appropriate signals to preserve this embryonic phenotype and serve as a matrix for ES cell adherence. Preferred feeder cells comprise primary embryonic fibroblasts that are established from tissue of day 13- day 14 embryos of virtually any mouse strain, that are maintained in culture, such as described by Abbondanzo et al.(1993) and are inhibited in growth by irradiation, such as described by Robertson (1987), or by the presence of an inhibitory concentration of LIF, such as described by Pease and Williams (1990).

The constructs in the host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence.

Following transformation of a suitable host and growth of the host to an appropriate cell density, the selected promoter is induced by appropriate means, such as temperature shift or chemical induction, and cells are cultivated for an additional period.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Microbial cells employed in the expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known by the skill artisan.

**Transgenic Animals**

The terms "transgenic animals" or "host animals" are used herein designate animals that have their genome genetically and artificially manipulated so as to include one of the nucleic acids according to the invention. Preferred animals are non-human mammals and include those belonging to a genus selected from *Mus* (e.g. mice), *Rattus* (e.g. rats) and *Oryctogalus* (e.g. rabbits) which have their genome artificially and genetically altered by the insertion of a nucleic acid according to the invention. In one embodiment, the invention encompasses non-human host mammals and animals comprising a recombinant vector of the invention or a BAP28 gene disrupted by homologous recombination with a knock out vector.

The transgenic animals of the invention all include within a plurality of their cells a cloned recombinant or synthetic DNA sequence, more specifically one of the purified or isolated nucleic acids comprising a *BAP28* coding sequence, a *BAP28* regulatory polynucleotide, a polynucleotide construct, or a DNA sequence encoding an antisense polynucleotide such as described in the present specification.

Generally, a transgenic animal according the present invention comprises any one of the polynucleotides, the recombinant vectors and the cell hosts described in the present invention. More particularly, the transgenic animals of the present invention can comprise any of the polynucleotides described in the "Genomic Sequences Of The *BAP28* Gene" section, the " *BAP28* cDNA

Sequences" section, the "Coding Regions" section, the "Polynucleotide constructs" section, the "Oligonucleotide Probes And Primers" section, the "Recombinant Vectors" section and the "Cell Hosts" section.

A further transgenic animals according to the invention contains in their somatic cells and/or in their germ line cells a polynucleotide comprising a biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably A1, A4, 16, A30, A31, A42, A50, A51, and A53, and the complements thereof.

In a first preferred embodiment, these transgenic animals may be good experimental models in order to study the diverse pathologies related to cell differentiation, in particular concerning the transgenic animals within the genome of which has been inserted one or several copies of a polynucleotide encoding a native BAP28 protein, or alternatively a mutant BAP28 protein.

In a second preferred embodiment, these transgenic animals may express a desired polypeptide of interest under the control of the regulatory polynucleotides of the *BAP28* gene, leading to good yields in the synthesis of this protein of interest, and eventually a tissue specific expression of this protein of interest.

The design of the transgenic animals of the invention may be made according to the conventional techniques well known from the one skilled in the art. For more details regarding the production of transgenic animals, and specifically transgenic mice, it may be referred to US Patents Nos 4,873,191, issued Oct. 10, 1989; 5,464,764 issued Nov 7, 1995; and 5,789,215, issued Aug 4, 1998; these documents being herein incorporated by reference to disclose methods producing transgenic mice.

Transgenic animals of the present invention are produced by the application of procedures which result in an animal with a genome that has incorporated exogenous genetic material. The procedure involves obtaining the genetic material, or a portion thereof, which encodes either a *BAP28* coding sequence, a *BAP28* regulatory polynucleotide or a DNA sequence encoding a *BAP28* antisense polynucleotide such as described in the present specification.

A recombinant polynucleotide of the invention is inserted into an embryonic or ES stem cell line. The insertion is preferably made using electroporation, such as described by Thomas et al.(1987). The cells subjected to electroporation are screened (e.g. by selection via selectable markers, by PCR or by Southern blot analysis) to find positive cells which have integrated the exogenous recombinant polynucleotide into their genome, preferably via an homologous recombination event. An illustrative positive-negative selection procedure that may be used according to the invention is described by Mansour et al.(1988).

Then, the positive cells are isolated, cloned and injected into 3.5 days old blastocysts from mice, such as described by Bradley (1987). The blastocysts are then inserted into a female host animal and allowed to grow to term.

Alternatively, the positive ES cells are brought into contact with embryos at the 2.5 days old 8-16 cell stage (morulae) such as described by Wood et al.(1993) or by Nagy et al.(1993), the ES cells being internalized to colonize extensively the blastocyst including the cells which will give rise to the germ line.

The offspring of the female host are tested to determine which animals are transgenic e.g. include the inserted exogenous DNA sequence and which are wild-type.

Thus, the present invention also concerns a transgenic animal containing a nucleic acid, a recombinant expression vector or a recombinant host cell according to the invention.

**Recombinant Cell Lines Derived From The Transgenic Animals Of The Invention.**

A further object of the invention consists of recombinant host cells obtained from a transgenic animal described herein. In one embodiment the invention encompasses cells derived from non-human host mammals and animals comprising a recombinant vector of the invention or a *BAP28* gene disrupted by homologous recombination with a knock out vector.

Recombinant cell lines may be established *in vitro* from cells obtained from any tissue of a transgenic animal according to the invention, for example by transfection of primary cell cultures with vectors expressing *onc*-genes such as SV40 large T antigen, as described by Chou (1989) and Shay et al.(1991).

**Methods for screening substances interacting with a BAP28 polypeptide**

For the purpose of the present invention, a ligand means a molecule, such as a protein, a peptide, an antibody or any synthetic chemical compound capable of binding to the BAP28 protein or one of its fragments or variants or to modulate the expression of the polynucleotide coding for BAP28 or a fragment or variant thereof.

In the ligand screening method according to the present invention, a biological sample or a defined molecule to be tested as a putative ligand of the BAP28 protein is brought into contact with the corresponding purified BAP28 protein, for example the corresponding purified recombinant BAP28 protein produced by a recombinant cell host as described hereinbefore, in order to form a complex between this protein and the putative ligand molecule to be tested.

As an illustrative example, to study the interaction of the BAP28 protein, or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5, or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino

acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, with drugs or small molecules, such as molecules generated through combinatorial chemistry approaches, the microdialysis coupled to HPLC method described by Wang et al. (1997) or the affinity capillary electrophoresis method described by Bush et al. (1997), the disclosures of which are incorporated by reference, can be used.

In further methods, peptides, drugs, fatty acids, lipoproteins, or small molecules which interact with the BAP28 protein, or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, may be identified using assays such as the following. The molecule to be tested for binding is labeled with a detectable label, such as a fluorescent , radioactive, or enzymatic tag and placed in contact with immobilized BAP28 protein, or a fragment thereof under conditions which permit specific binding to occur. After removal of non-specifically bound molecules, bound molecules are detected using appropriate means.

Another object of the present invention consists of methods and kits for the screening of candidate substances that interact with BAP28 polypeptide.

The present invention pertains to methods for screening substances of interest that interact with a BAP28 protein or one fragment or variant thereof. By their capacity to bind covalently or non-covalently to a BAP28 protein or to a fragment or variant thereof, these substances or molecules may be advantageously used both *in vitro* and *in vivo*.

*In vitro*, said interacting molecules may be used as detection means in order to identify the presence of a BAP28 protein in a sample, preferably a biological sample.

A method for the screening of a candidate substance comprises the following steps :

a) providing a polypeptide consisting of a BAP28 protein or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at

the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, or a variant thereof;

b) obtaining a candidate substance;

c) bringing into contact said polypeptide with said candidate substance;

d) detecting the complexes formed between said polypeptide and said candidate substance.

The invention further concerns a kit for the screening of a candidate substance interacting with the BAP28 polypeptide, wherein said kit comprises :

a) a BAP28 protein having an amino acid sequence selected from the group consisting of the amino acid sequences of SEQ ID No 5 or a peptide fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, or a variant thereof ;

b) in some embodiments, the kit may also comprise means useful to detect the complex formed between the BAP28 protein or a peptide fragment or a variant thereof and the candidate substance.

In a preferred embodiment of the kit described above, the detection means consist in monoclonal or polyclonal antibodies directed against the BAP28 protein or a peptide fragment or a variant thereof.

Various candidate substances or molecules can be assayed for interaction with a BAP28 polypeptide. These substances or molecules include, without being limited to, natural or synthetic organic compounds or molecules of biological origin such as polypeptides. When the candidate substance or molecule consists of a polypeptide, this polypeptide may be the resulting expression product of a phage clone belonging to a phage-based random peptide library, or alternatively the polypeptide may be the resulting expression product of a cDNA library cloned in a vector suitable for performing a two-hybrid screening assay.

The invention also pertains to kits useful for performing the hereinbefore described screening method. Preferably, such kits comprise a BAP28 polypeptide or a fragment or a variant thereof, and, in some embodiments, means useful to detect the complex formed between the BAP28 polypeptide or its fragment or variant and the candidate substance. In a preferred embodiment the detection means consist in monoclonal or polyclonal antibodies directed against the corresponding BAP28 polypeptide or a fragment or a variant thereof.

**A. Candidate ligands obtained from random peptide libraries**

In a particular embodiment of the screening method, the putative ligand is the expression product of a DNA insert contained in a phage vector (Parmley and Smith, 1988). Specifically, random peptide phages libraries are used. The random DNA inserts encode for peptides of 8 to 20 amino acids in length (Oldenburg K.R. et al., 1992; Valadon P., et al., 1996; Lucas A.H., 1994; Westerink M.A.J., 1995; Felici F. et al., 1991). According to this particular embodiment, the recombinant phages expressing a protein that binds to the immobilized BAP28 protein is retained and the complex formed between the BAP28 protein and the recombinant phage may be subsequently immunoprecipitated by a polyclonal or a monoclonal antibody directed against the BAP28 protein.

Once the ligand library in recombinant phages has been constructed, the phage population is brought into contact with the immobilized BAP28 protein. Then the preparation of complexes is washed in order to remove the non-specifically bound recombinant phages. The phages that bind specifically to the BAP28 protein are then eluted by a buffer (acid pH) or immunoprecipitated by the monoclonal antibody produced by the hybridoma anti-BAP28, and this phage population is subsequently amplified by an over-infection of bacteria (for example E. coli). The selection step may be repeated several times, preferably 2-4 times, in order to select the more specific recombinant phage clones. The last step consists in characterizing the peptide produced by the selected recombinant phage clones either by expression in infected bacteria and isolation, expressing the phage insert in another host-vector system, or sequencing the insert contained in the selected recombinant phages.

**B. Candidate ligands obtained by competition experiments.**

Alternatively, peptides, drugs or small molecules which bind to the BAP28 protein, or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, may be identified in competition experiments. In such assays, the BAP28 protein, or a fragment thereof, is immobilized to a surface, such as a plastic plate. Increasing amounts of the peptides, drugs or small molecules are placed in contact with the immobilized BAP28 protein, or a fragment thereof, in the presence of a detectable labeled known BAP28 protein ligand. For example, the BAP28 ligand may be detectably labeled with a fluorescent, radioactive, or enzymatic tag. The ability of the test molecule to bind the BAP28

protein, or a fragment thereof, is determined by measuring the amount of detectably labeled known ligand bound in the presence of the test molecule. A decrease in the amount of known ligand bound to the BAP28 protein, or a fragment thereof, when the test molecule is present indicated that the test molecule is able to bind to the BAP28 protein, or a fragment thereof.

**C. Candidate ligands obtained by affinity chromatography.**

Proteins or other molecules interacting with the BAP28 protein, or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, can also be found using affinity columns which contain the BAP28 protein, or a fragment thereof. The BAP28 protein, or a fragment thereof, may be attached to the column using conventional techniques including chemical coupling to a suitable column matrix such as agarose, Affi Gel® , or other matrices familiar to those of skill in art. In some embodiments of this method, the affinity column contains chimeric proteins in which the BAP28 protein, or a fragment thereof, is fused to glutathion S transferase (GST). A mixture of cellular proteins or pool of expressed proteins as described above is applied to the affinity column. Proteins or other molecules interacting with the BAP28 protein, or a fragment thereof, attached to the column can then be isolated and analyzed on 2-D electrophoresis gel as described in Ramunsen et al. (1997), the disclosure of which is incorporated by reference. Alternatively, the proteins retained on the affinity column can be purified by electrophoresis based methods and sequenced. The same method can be used to isolate antibodies, to screen phage display products, or to screen phage display human antibodies.

**D. Candidate ligands obtained by optical biosensor methods**

Proteins interacting with the BAP28 protein, or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, can also be screened by using an Optical Biosensor as described in Edwards and Leatherbarrow (1997) and also in Szabo et al. (1995), the disclosure of which is incorporated by reference. This technique

permits the detection of interactions between molecules in real time, without the need of labeled molecules. This technique is based on the surface plasmon resonance (SPR) phenomenon. Briefly, the candidate ligand molecule to be tested is attached to a surface (such as a carboxymethyl dextran matrix). A light beam is directed towards the side of the surface that does not contain the sample to be tested and is reflected by said surface. The SPR phenomenon causes a decrease in the intensity of the reflected light with a specific association of angle and wavelength. The binding of candidate ligand molecules cause a change in the refraction index on the surface, which change is detected as a change in the SPR signal. For screening of candidate ligand molecules or substances that are able to interact with the BAP28 protein, or a fragment thereof, the BAP28 protein, or a fragment thereof, is immobilized onto a surface. This surface consists of one side of a cell through which flows the candidate molecule to be assayed. The binding of the candidate molecule on the BAP28 protein, or a fragment thereof, is detected as a change of the SPR signal. The candidate molecules tested may be proteins, peptides, carbohydrates, lipids, or small molecules generated by combinatorial chemistry. This technique may also be performed by immobilizing eukaryotic or prokaryotic cells or lipid vesicles exhibiting an endogenous or a recombinantly expressed BAP28 protein at their surface.

The main advantage of the method is that it allows the determination of the association rate between the BAP28 protein and molecules interacting with the BAP28 protein. It is thus possible to select specifically ligand molecules interacting with the BAP28 protein, or a fragment thereof, through strong or conversely weak association constants.

**E. Candidate ligands obtained through a two-hybrid screening assay.**

The yeast two-hybrid system is designed to study protein-protein interactions *in vivo* (Fields and Song, 1989), and relies upon the fusion of a bait protein to the DNA binding domain of the yeast Gal4 protein. This technique is also described in the US Patent No US 5,667,973 and the US Patent No 5,283,173 (Fields et al.) the technical teachings of both patents being herein incorporated by reference.

The general procedure of library screening by the two-hybrid assay may be performed as described by Harper et al. (1993) or as described by Cho et al. (1998) or also Fromont-Racine et al. (1997).

The bait protein or polypeptide consists of a BAP28 polypeptide or a fragment comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at

the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5, or a variant thereof.

More precisely, the nucleotide sequence encoding the BAP28 polypeptide or a fragment or variant thereof is fused to a polynucleotide encoding the DNA binding domain of the GAL4 protein, the fused nucleotide sequence being inserted in a suitable expression vector, for example pAS2 or pM3.

Then, a human cDNA library is constructed in a specially designed vector, such that the human cDNA insert is fused to a nucleotide sequence in the vector that encodes the transcriptional domain of the GAL4 protein. Preferably, the vector used is the pACT vector. The polypeptides encoded by the nucleotide inserts of the human cDNA library are termed "pray" polypeptides.

A third vector contains a detectable marker gene, such as beta galactosidase gene or CAT gene that is placed under the control of a regulation sequence that is responsive to the binding of a complete Gal4 protein containing both the transcriptional activation domain and the DNA binding domain. For example, the vector pG5EC may be used.

Two different yeast strains are also used. As an illustrative but non limiting example the two different yeast strains may be the followings :

- Y190, the phenotype of which is (*MATa, Leu2-3, 112 ura3-12, trp1-901, his3-D200, ade2-101, gal4Dgal180D URA3 GAL-LacZ, LYS GAL-HIS3, cyh*$^r$);
- Y187, the phenotype of which is (*MATa gal4 gal80 his3 trp1-901 ade2-101 ura3-52 leu2-3, -112 URA3 GAL-lacZmet*$^-$), which is the opposite mating type of Y190.

Briefly, 20 µg of pAS2/*BAP28* and 20 µg of pACT-cDNA library are co-transformed into yeast strain Y190. The transformants are selected for growth on minimal media lacking histidine, leucine and tryptophan, but containing the histidine synthesis inhibitor 3-AT (50 mM). Positive colonies are screened for beta galactosidase by filter lift assay. The double positive colonies (*His*$^+$, *beta-gal*$^+$) are then grown on plates lacking histidine, leucine, but containing tryptophan and cycloheximide (10 mg/ml) to select for loss of pAS2/*BAP28* plasmids bu retention of pACT-cDNA library plasmids. The resulting Y190 strains are mated with Y187 strains expressing BAP28 or non-related control proteins; such as cyclophilin B, lamin, or SNF1, as *Gal4* fusions as described by Harper et al. (1993) and by Bram et al. (Bram RJ et al., 1993), and screened for beta galactosidase by filter lift assay. Yeast clones that are *beta gal-* after mating with the control *Gal4* fusions are considered false positives.

In another embodiment of the two-hybrid method according to the invention, interaction between the BAP28 or a fragment or variant thereof with cellular proteins may be assessed using the Matchmaker Two Hybrid System 2 (Catalog No K1604-1, Clontech). As described in the manual accompanying the Matchmaker Two Hybrid System 2 (Catalog No K1604-1, Clontech), the disclosure of which is incorporated herein by reference, nucleic acids encoding the BAP28 protein or a portion thereof, are inserted into an expression vector such that they are in frame with DNA encoding the DNA

binding domain of the yeast transcriptional activator GAL4. A desired cDNA, preferably human cDNA, is inserted into a second expression vector such that they are in frame with DNA encoding the activation domain of GAL4. The two expression plasmids are transformed into yeast and the yeast are plated on selection medium which selects for expression of selectable markers on each of the expression vectors as well as GAL4 dependent expression of the HIS3 gene. Transformants capable of growing on medium lacking histidine are screened for GAL4 dependent lacZ expression. Those cells which are positive in both the histidine selection and the lacZ assay contain interaction between BAP28 and the protein or peptide encoded by the initially selected cDNA insert.

**Methods For Screening Substances Modulating The Activity Of The BAP28 protein**

The invention also concerns a method for screening new agents, or candidate substances which modulate the activity of the BAP28 protein or a fragment thereof. Preferably, the BAP28 protein or a fragment thereof is a polypeptide code comprising a contiguous span of at least 6 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5. Preferably, the candidate substance is mixed with the BAP28 protein and the activity of the BAP28 protein is measured. Candidate substances include, without being limited to, natural or synthetic organic compounds or molecules of biological origin such as polypeptides.

**Method For Screening Substances Interacting With The Regulatory Sequences Of The *BAP28* Gene**

The present invention also concerns a method for screening substances or molecules that are able to interact with the regulatory sequences of the *BAP28* gene, such as for example promoter or enhancer sequences.

Nucleic acids encoding proteins which are able to interact with the regulatory sequences of the *BAP28* gene, more particularly a nucleotide sequence selected from the group consisting of the polynucleotides of the 5' and 3' regulatory region or a fragment or variant thereof, and preferably a variant comprising one of the biallelic markers of the invention, may be identified by using a one-hybrid system, such as that described in the booklet enclosed in the Matchmaker One-Hybrid System kit from Clontech (Catalog Ref. n° K1603-1), the technical teachings of which are herein incorporated by reference. Briefly, the target nucleotide sequence is cloned upstream of a selectable reporter sequence and the resulting DNA construct is integrated in the yeast genome (*Saccharomyces cerevisiae*). The yeast cells containing the reporter sequence in their genome are then transformed with a library consisting of fusion molecules between cDNAs encoding candidate proteins for binding onto the regulatory sequences of the *BAP28* gene and sequences encoding the

activator domain of a yeast transcription factor such as GAL4. The recombinant yeast cells are plated in a culture broth for selecting cells expressing the reporter sequence. The recombinant yeast cells thus selected contain a fusion protein that is able to bind onto the target regulatory sequence of the *BAP28* gene. Then, the cDNAs encoding the fusion proteins are sequenced and may be cloned into expression or transcription vectors *in vitro*. The binding of the encoded polypeptides to the target regulatory sequences of the *BAP28* gene may be confirmed by techniques familiar to the one skilled in the art, such as gel retardation assays or DNAse protection assays.

Gel retardation assays may also be performed independently in order to screen candidate molecules that are able to interact with the regulatory sequences of the *BAP28* gene, such as described by Fried and Crothers (1981), Garner and Revzin (1981) and Dent and Latchman (1993), the teachings of these publications being herein incorporated by reference. These techniques are based on the principle according to which a DNA fragment which is bound to a protein migrates slower than the same unbound DNA fragment. Briefly, the target nucleotide sequence is labeled. Then the labeled target nucleotide sequence is brought into contact with either a total nuclear extract from cells containing transcription factors, or with different candidate molecules to be tested. The interaction between the target regulatory sequence of the *BAP28* gene and the candidate molecule or the transcription factor is detected after gel or capillary electrophoresis through a retardation in the migration.

**Method For Screening Ligands That Modulate The Expression Of The BAP28 Protein**

Another subject of the present invention is a method for screening molecules that modulate the expression of the BAP28 protein. Such a screening method comprises the steps of:

a) cultivating a prokaryotic or an eukaryotic cell that has been transfected with a nucleotide sequence encoding the BAP28 protein or a variant or a fragment thereof, placed under the control of its own promoter;

b) bringing into contact the cultivated cell with a molecule to be tested;

c) quantifying the expression of the BAP28 protein or a variant or a fragment thereof.

Using DNA recombination techniques well known by the one skill in the art, the BAP28 protein encoding DNA sequence is inserted into an expression vector, downstream from its promoter sequence. As an illustrative example, the promoter sequence of the *BAP28* gene is contained in the nucleic acid of the 5' regulatory region.

The quantification of the expression of the BAP28 protein may be realized either at the mRNA level or at the protein level. In the latter case, polyclonal or monoclonal antibodies may be used to quantify the amounts of the BAP28 protein that have been produced, for example in an ELISA or a RIA assay.

In a preferred embodiment, the quantification of the *BAP28* mRNA is realized by a quantitative PCR amplification of the cDNA obtained by a reverse transcription of the total mRNA of the cultivated *BAP28* -transfected host cell, using a pair of primers specific for *BAP28*.

The present invention also concerns a method for screening substances or molecules that are able to increase, or in contrast to decrease, the level of expression of the *BAP28* gene. Such a method may allow the one skilled in the art to select substances exerting a regulating effect on the expression level of the *BAP28* gene and which may be useful as active ingredients included in pharmaceutical compositions for treating patients suffering from prostate cancer.

Thus, is also part of the present invention a method for screening of a candidate substance or molecule that modulated the expression of the *BAP28* gene, this method comprises the following steps:

- providing a recombinant cell host containing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence of the 5' regulatory region or a biologically active fragment or variant thereof located upstream a polynucleotide encoding a detectable protein;

- obtaining a candidate substance; and

- determining the ability of the candidate substance to modulate the expression levels of the polynucleotide encoding the detectable protein.

In a further embodiment, the nucleic acid comprising the nucleotide sequence of the 5' regulatory region or a biologically active fragment or variant thereof also includes a 5'UTR region of the *BAP28* cDNA of SEQ ID No 2 or 3, or one of its biologically active fragments or variants thereof.

Among the preferred polynucleotides encoding a detectable protein, there may be cited polynucleotides encoding beta galactosidase, green fluorescent protein (GFP) and chloramphenicol acetyl transferase (CAT). In some embodiments, the detectable protein can be BAP28 or a fragment thereof.

The invention also pertains to kits useful for performing the hereinbefore described screening method. Preferably, such kits comprise a recombinant vector that allows the expression of a nucleotide sequence of the 5' regulatory region or a biologically active fragment or variant thereof located upstream and operably linked to a polynucleotide encoding a detectable protein or the BAP28 protein or a fragment or a variant thereof.

In another embodiment of a method for the screening of a candidate substance or molecule that modulates the expression of the *BAP28* gene, wherein said method comprises the following steps:

a) providing a recombinant host cell containing a nucleic acid, wherein said nucleic acid comprises a 5'UTR sequence of the *BAP28* cDNA of SEQ ID No 2 or 3, or one of its biologically active fragments or variants, the 5'UTR sequence or its biologically active fragment or variant being operably linked to a polynucleotide encoding a detectable protein;

b) obtaining a candidate substance; and

c) determining the ability of the candidate substance to modulate the expression levels of the polynucleotide encoding the detectable protein.

In a specific embodiment of the above screening method, the nucleic acid that comprises a nucleotide sequence selected from the group consisting of the 5'UTR sequence of the *BAP28* cDNA of SEQ ID No 2 or 3 or one of its biologically active fragments or variants, includes a promoter sequence which is endogenous with respect to the *BAP28* 5'UTR sequence.

In another specific embodiment of the above screening method, the nucleic acid that comprises a nucleotide sequence selected from the group consisting of the 5'UTR sequence of the *BAP28* cDNA of SEQ ID No 2 or 3 or one of its biologically active fragments or variants, includes a promoter sequence which is exogenous with respect to the *BAP28* 5'UTR sequence defined therein.

In a further preferred embodiment, the nucleic acid comprising the 5'-UTR sequence of the *BAP28* cDNA or SEQ ID No 2 or 3 or the biologically active fragments thereof includes a biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53, or the complements thereof.

The invention further deals with a kit for the screening of a candidate substance modulating the expression of the *BAP28* gene, wherein said kit comprises a recombinant vector that comprises a nucleic acid including a 5'UTR sequence of the *BAP28* cDNA of SEQ ID No 2 or 3, or one of their biologically active fragments or variants, the 5'UTR sequence or its biologically active fragment or variant being operably linked to a polynucleotide encoding a detectable protein.

For the design of suitable recombinant vectors useful for performing the screening methods described above, it will be referred to the section of the present specification wherein the preferred recombinant vectors of the invention are detailed.

Expression levels and patterns of *BAP28* may be analyzed by solution hybridization with long probes as described in International Patent Application No WO 97/05277, the entire contents of which are incorporated herein by reference. Briefly, the *BAP28* cDNA or the *BAP28* genomic DNA described above, or fragments thereof, is inserted at a cloning site immediately downstream of a bacteriophage (T3, T7 or SP6) RNA polymerase promoter to produce antisense RNA. Preferably, the *BAP28* insert comprises at least 100 or more consecutive nucleotides of the genomic DNA sequence or the cDNA sequences. The plasmid is linearized and transcribed in the presence of ribonucleotides comprising modified ribonucleotides (i.e. biotin-UTP and DIG-UTP). An excess of this doubly labeled RNA is hybridized in solution with mRNA isolated from cells or tissues of interest. The hybridizations are performed under standard stringent conditions (40-50°C for 16 hours in an 80% formamide, 0. 4 M NaCl buffer, pH 7-8). The unhybridized probe is removed by digestion with ribonucleases specific for single-stranded RNA (i.e. RNases CL3, T1, Phy M, U2 or A). The presence of the biotin-UTP modification enables capture of the hybrid on a microtitration

plate coated with streptavidin. The presence of the DIG modification enables the hybrid to be detected and quantified by ELISA using an anti-DIG antibody coupled to alkaline phosphatase.

Quantitative analysis of *BAP28* gene expression may also be performed using arrays. As used herein, the term array means a one dimensional, two dimensional, or multidimensional arrangement of a plurality of nucleic acids of sufficient length to permit specific detection of expression of mRNAs capable of hybridizing thereto. For example, the arrays may contain a plurality of nucleic acids derived from genes whose expression levels are to be assessed. The arrays may include the *BAP28* genomic DNA, the *BAP28* cDNA sequences or the sequences complementary thereto or fragments thereof, particularly those comprising at least one of the biallelic markers according the present invention, preferably at least one of the biallelic markers A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably at least one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53. Preferably, the fragments are at least 15 nucleotides in length. In other embodiments, the fragments are at least 25 nucleotides in length. In some embodiments, the fragments are at least 50 nucleotides in length. More preferably, the fragments are at least 100 nucleotides in length. In another preferred embodiment, the fragments are more than 100 nucleotides in length. In some embodiments the fragments may be more than 500 nucleotides in length.

For example, quantitative analysis of *BAP28* gene expression may be performed with a complementary DNA microarray as described by Schena et al.(1995 and 1996). Full length *BAP28* cDNAs or fragments thereof are amplified by PCR and arrayed from a 96-well microtiter plate onto silylated microscope slides using high-speed robotics. Printed arrays are incubated in a humid chamber to allow rehydration of the array elements and rinsed, once in 0.2% SDS for 1 min, twice in water for 1 min and once for 5 min in sodium borohydride solution. The arrays are submerged in water for 2 min at 95°C, transferred into 0.2% SDS for 1 min, rinsed twice with water, air dried and stored in the dark at 25°C.

Cell or tissue mRNA is isolated or commercially obtained and probes are prepared by a single round of reverse transcription. Probes are hybridized to 1 $cm^2$ microarrays under a 14 x 14 mm glass coverslip for 6-12 hours at 60°C. Arrays are washed for 5 min at 25°C in low stringency wash buffer (1 x SSC/0.2% SDS), then for 10 min at room temperature in high stringency wash buffer (0.1 x SSC/0.2% SDS). Arrays are scanned in 0.1 x SSC using a fluorescence laser scanning device fitted with a custom filter set. Accurate differential expression measurements are obtained by taking the average of the ratios of two independent hybridizations.

Quantitative analysis of *BAP28* gene expression may also be performed with full length *BAP28* cDNAs or fragments thereof in complementary DNA arrays as described by Pietu et al.(1996). The full length *BAP28* cDNA or fragments thereof is PCR amplified and spotted on membranes. Then, mRNAs originating from various tissues or cells are labeled with radioactive nucleotides. After hybridization and washing in controlled conditions, the hybridized mRNAs are

detected by phospho-imaging or autoradiography. Duplicate experiments are performed and a quantitative analysis of differentially expressed mRNAs is then performed.

Alternatively, expression analysis using the *BAP28* genomic DNA, the *BAP28* cDNA, or fragments thereof can be done through high density nucleotide arrays as described by Lockhart et al.(1996) and Sosnowsky et al.(1997). Oligonucleotides of 15-50 nucleotides from the sequences of the *BAP28* genomic DNA, the *BAP28* cDNA sequences particularly those comprising at least one of biallelic markers according the present invention, preferably at least one biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably at least one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53, or the sequences complementary thereto, are synthesized directly on the chip (Lockhart et al., supra) or synthesized and then addressed to the chip (Sosnowski et al., supra). Preferably, the oligonucleotides are about 20 nucleotides in length.

BAP28 cDNA probes labeled with an appropriate compound, such as biotin, digoxigenin or fluorescent dye, are synthesized from the appropriate mRNA population and then randomly fragmented to an average size of 50 to 100 nucleotides. The said probes are then hybridized to the chip. After washing as described in Lockhart et al., supra and application of different electric fields (Sosnowsky et al., 1997), the dyes or labeling compounds are detected and quantified. Duplicate hybridizations are performed. Comparative analysis of the intensity of the signal originating from cDNA probes on the same target oligonucleotide in different cDNA samples indicates a differential expression of BAP28 mRNA.

**Computer-Related Embodiments**

As used herein the term "nucleic acid codes of the invention" encompass the nucleotide sequences comprising, consisting essentially of, or consisting of any one of the following:

a) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 1, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 1: 1-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662;

b) a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, or 500 nucleotides of SEQ ID No 1 or the complement thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 20, 30, 40 or 50 nucleotides selected from the group consisting of the following nucleotide positions of SEQ ID No 1: 4997-5076, 5371-5544, 6121-6337, 9877-10018, 11522-11623, 12521-12661, 13453-13664, 13824-13957, 15376-15478, 16855-16965, 17378-17495, 18535-18642, 21446-21541, 21999-22087, 23036-23247, 23546-23667, 24270-24461, 26287-26470, 26611-26747, 28068-28260, 32540-32709, 33112-33270, 34586-34828, 35156-35287, 36660-36763, 36934-37077, 37803-37921, 38017-38138, 40365-40493, 42618-42848, 43452-43578, 44836-44999, 48223-48269, and 49656-49779;

c) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 1 or the complements thereof, wherein said contiguous span comprises at least one BAP28-related biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53;

d) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of nucleotide positions 1 to 4995 of SEQ ID No 2 or 3;

e) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of nucleotide positions 1 to 2033, 2160 to 2348 and 2676 to 4995 of SEQ ID No 2 or 3;

f) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 1-3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of any one of the following ranges of nucleotide positions of:

(1) SEQ ID No 1: 1-2500, 2501-5000, 5001-7500, 7501-10000, 10001-12500, 12501-15000, 15001-17500, 17501-20000, 20001-22500, 22501-25000, 25001-27500, 27501-30000, 30001-32500, 32501-35000, 35001-37500, 37501-40000, 40001-42500, 42501-45000, 45001-47500, 47501-50000, 50001-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662;

(2) SEQ ID No 2: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, and 6501 to 6782; and,

(3) SEQ ID No 3: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, 6501 to 7000, 7001 to 7500, 7501 to 7932; and

g) a nucleotide sequence selected from the group consisting of SEQ ID Nos 4, and 9-13; and,

h) a nucleotide sequence complementary to any one of the preceding nucleotide sequences.

The "nucleic acid codes of the invention" further encompass nucleotide sequences homologous to:

a) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 1, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 1: 1-50357, 50499-50963, 51257-

52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662;

b) a contiguous span of at least 12, 15, 18, 20, 25, 30, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, or 500 nucleotides of SEQ ID No 1 or the complement thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 20, 30, 40 or 50 nucleotides selected from the group consisting of the following nucleotide positions of SEQ ID No 1: 4997-5076, 5371-5544, 6121-6337, 9877-10018, 11522-11623, 12521-12661, 13453-13664, 13824-13957, 15376-15478, 16855-16965, 17378-17495, 18535-18642, 21446-21541, 21999-22087, 23036-23247, 23546-23667, 24270-24461, 26287-26470, 26611-26747, 28068-28260, 32540-32709, 33112-33270, 34586-34828, 35156-35287, 36660-36763, 36934-37077, 37803-37921, 38017-38138, 40365-40493, 42618-42848, 43452-43578, 44836-44999, 48223-48269, and 49656-49779;

c) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 1 or the complements thereof, wherein said contiguous span comprises at least one BAP28-related biallelic marker selected from the group consisting of A1 to A58, preferably A1 to A27, A34, A37 to A41, A43 to A49, A52, and A54 to A58, more preferably one of the biallelic markers A1, A4, 16, A30, A31, A42, A50, A51, and A53;

d) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of nucleotide positions 1 to 4995 of SEQ ID No 2 or 3;

e) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 2 and 3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of nucleotide positions 1 to 2033, 2160 to 2348 and 2676 to 4995 of SEQ ID No 2 or 3;

f) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of a nucleic acid sequence selected from the group consisting of SEQ ID Nos 1-3 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of any one of the following ranges of nucleotide positions of:

(1) SEQ ID No 1: 1-2500, 2501-5000, 5001-7500, 7501-10000, 10001-12500, 12501-15000, 15001-17500, 17501-20000, 20001-22500, 22501-25000, 25001-27500, 27501-30000, 30001-32500, 32501-35000, 35001-37500, 37501-40000, 40001-42500, 42501-45000, 45001-47500, 47501-50000, 50001-50357, 50499-50963, 51257-52147, 52299-53234, 53394-53553, 53689-53837, 53943-54028, 54198-54740, 54896-55753, 55913-57385, 57495-58503, 58828-85946, 59355-85946, 86169-91228, and/or 91852 to 97662;

(2) SEQ ID No 2: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, and 6501 to 6782; and,

(3) SEQ ID No 3: 1 to 500, 501 to 1000, 1001 to 1500, 1501 to 2000, 2001 to 2500, 2501 to 3000, 3001 to 3500, 3501 to 4000, 4001 to 4500, 4501 to 4995, 5000 to 5500, 5501 to 6000, 6001 to 6500, 6501 to 7000, 7001 to 7500, 7501 to 7932; and

g) a nueclotide sequence selected from the group consisting of SEQ ID Nos 4, and 9-13; and,

h) a nucleotide sequence complementary to any one of the preceding nucleotide sequences.

Homologous sequences refer to a sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, or 75% homology to these contiguous spans. Homology may be determined using any method described herein, including BLAST2N with the default parameters or with any modified parameters. Homologous sequences also may include RNA sequences in which uridines replace the thymines in the nucleic acid codes of the invention. It will be appreciated that the nucleic acid codes of the invention can be represented in the traditional single character format (See the inside back cover of Stryer, Lubert. *Biochemistry*, 3rd edition. W. H Freeman & Co., New York.) or in any other format or code which records the identity of the nucleotides in a sequence.

As used herein the term "polypeptide codes of the invention" encompass the polypeptide sequences comprising a contiguous span of at least 6, 8, 10, 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 5, wherein said contiguous span includes either at least 1, 2, 3, 5 or 10 of the amino acid positions selected from the group consisting of 1 to 1629 of the SEQ ID No 5 or an amino acid selected from the group consisting of an asparagine at the amino acid position 1694 of SEQ ID No 5, a valine at the amino acid position 1854 of SEQ ID No 5, an asparagine at the amino acid position 1967 of SEQ ID No 5, a glutamic acid at the amino acid position 2017 of SEQ ID No 5, and an alanine at the amino acid position 2050 of SEQ ID No 5. It will be appreciated that the polypeptide codes of the invention can be represented in the traditional single character format or three letter format (See the inside back cover of Stryer, Lubert. Biochemistry, 3rd edition. W. H Freeman & Co., New York.) or in any other format or code which records the identity of the polypeptides in a sequence.

It will be appreciated by those skilled in the art that the nucleic acid codes of the invention and polypeptide codes of the invention can be stored, recorded, and manipulated on any medium which can be read and accessed by a computer. As used herein, the words "recorded" and "stored" refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any of the presently known methods for recording information on a computer readable medium to generate manufactures comprising one or more of the nucleic acid codes of the invention, or one or more of the polypeptide codes of the invention. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, 20, 25, 30, or 50 nucleic acid codes of the invention. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, 20, 25, 30, or 50 polypeptide codes of the invention.

Computer readable media include magnetically readable media, optically readable media, electronically readable media and magnetic/optical media. For example, the computer readable media may be a hard disk, a floppy disk, a magnetic tape, CD-ROM, Digital Versatile Disk (DVD), Random Access Memory (RAM), or Read Only Memory (ROM) as well as other types of other media known to those skilled in the art.

Embodiments of the present invention include systems, particularly computer systems which store and manipulate the sequence information described herein. One example of a computer system 100 is illustrated in block diagram form in Figure 7. As used herein, "a computer system" refers to the hardware components, software components, and data storage components used to analyze the nucleotide sequences of the nucleic acid codes of the invention or the amino acid sequences of the polypeptide codes of the invention. In one embodiment, the computer system 100 is a Sun Enterprise 1000 server (Sun Microsystems, Palo Alto, CA). The computer system 100 preferably includes a processor for processing, accessing and manipulating the sequence data. The processor 105 can be any well-known type of central processing unit, such as the Pentium III from Intel Corporation, or similar processor from Sun, Motorola, Compaq or International Business Machines.

Preferably, the computer system 100 is a general purpose system that comprises the processor 105 and one or more internal data storage components 110 for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

In one particular embodiment, the computer system 100 includes a processor 105 connected to a bus which is connected to a main memory 115 (preferably implemented as RAM) and one or more internal data storage devices 110, such as a hard drive and/or other computer readable media having data recorded thereon. In some embodiments, the computer system 100 further includes one or more data retrieving device 118 for reading the data stored on the internal data storage devices 110.

The data retrieving device 118 may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, etc. In some embodiments, the internal data storage device 110 is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system 100 may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device.

The computer system 100 includes a display 120 which is used to display output to a computer user. It should also be noted that the computer system 100 can be linked to other computer systems 125a-c in a network or wide area network to provide centralized access to the computer system 100.

Software for accessing and processing the nucleotide sequences of the nucleic acid codes of the invention or the amino acid sequences of the polypeptide codes of the invention (such as search tools, compare tools, and modeling tools etc.) may reside in main memory 115 during execution.

In some embodiments, the computer system 100 may further comprise a sequence comparer for comparing the above-described nucleic acid codes of the invention or the polypeptide codes of the invention stored on a computer readable medium to reference nucleotide or polypeptide sequences stored on a computer readable medium. A "sequence comparer" refers to one or more programs which are implemented on the computer system 100 to compare a nucleotide or polypeptide sequence with other nucleotide or polypeptide sequences and/or compounds including but not limited to peptides, peptidomimetics, and chemicals stored within the data storage means. For example, the sequence comparer may compare the nucleotide sequences of nucleic acid codes of the invention or the amino acid sequences of the polypeptide codes of the invention stored on a computer readable medium to reference sequences stored on a computer readable medium to identify homologies, motifs implicated in biological function, or structural motifs. The various sequence comparer programs identified elsewhere in this patent specification are particularly contemplated for use in this aspect of the invention.

Figure 8 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database. The database of sequences can be a private database stored within the computer system 100, or a public database such as GENBANK, PIR OR SWISSPROT that is available through the Internet.

The process 200 begins at a start state 201 and then moves to a state 202 wherein the new sequence to be compared is stored to a memory in a computer system 100. As discussed above, the memory could be any type of memory, including RAM or an internal storage device.

The process 200 then moves to a state 204 wherein a database of sequences is opened for analysis and comparison. The process 200 then moves to a state 206 wherein the first sequence stored in the database is read into a memory on the computer. A comparison is then performed at a state 210 to determine if the first sequence is the same as the second sequence. It is important to note that this step is not limited to performing an exact comparison between the new sequence and the first sequence in the database. Well-known methods are known to those of skill in the art for comparing two nucleotide or protein sequences, even if they are not identical. For example, gaps can be introduced into one sequence in order to raise the homology level between the two tested sequences. The parameters that control whether gaps or other features are introduced into a sequence during comparison are normally entered by the user of the computer system.

Once a comparison of the two sequences has been performed at the state 210, a determination is made at a decision state 210 whether the two sequences are the same. Of course, the term "same" is not limited to sequences that are absolutely identical. Sequences that are within the homology parameters entered by the user will be marked as "same" in the process 200.

If a determination is made that the two sequences are the same, the process 200 moves to a state 214 wherein the name of the sequence from the database is displayed to the user. This state notifies the user that the sequence with the displayed name fulfills the homology constraints that were entered.

Once the name of the stored sequence is displayed to the user, the process 200 moves to a decision state 218 wherein a determination is made whether more sequences exist in the database. If no more sequences exist in the database, then the process 200 terminates at an end state 220. However, if more sequences do exist in the database, then the process 200 moves to a state 224 wherein a pointer is moved to the next sequence in the database so that it can be compared to the new sequence. In this manner, the new sequence is aligned and compared with every sequence in the database.

It should be noted that if a determination had been made at the decision state 212 that the sequences were not homologous, then the process 200 would move immediately to the decision state 218 in order to determine if any other sequences were available in the database for comparison.

Accordingly, one aspect of the present invention is a computer system comprising a processor, a data storage device having stored thereon a nucleic acid code of the invention or a polypeptide code of the invention, a data storage device having retrievably stored thereon reference nucleotide sequences or polypeptide sequences to be compared to the nucleic acid code of the invention or polypeptide code of the invention and a sequence comparer for conducting the comparison. The sequence comparer may indicate a homology level between the sequences compared or identify structural motifs in the nucleic acid code of the invention and polypeptide codes of the invention or it may identify structural motifs in sequences which are compared to these nucleic acid codes and polypeptide codes. In some embodiments, the data storage device may have stored thereon the sequences of at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of the invention or polypeptide codes of the invention.

Another aspect of the present invention is a method for determining the level of homology between a nucleic acid code of the invention and a reference nucleotide sequence, comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through the use of a computer program which determines homology levels and determining homology between the nucleic acid code and the reference nucleotide sequence with the computer program. The computer program may be any of a number of computer programs for determining homology levels, including those specifically enumerated herein, including BLAST2N with the default parameters or with any modified parameters. The method may be implemented using the computer systems described above. The method may also be performed by reading 2, 5, 10, 15, 20, 25, 30, or 50 of the above described nucleic acid codes of the invention through the use of the computer program and determining homology between the nucleic acid codes and reference nucleotide sequences.

Figure 9 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous. The process 250 begins at a start state 252 and then moves to a state 254 wherein a first sequence to be compared is stored to a memory. The second sequence to be compared is then stored to a memory at a state 256. The process 250 then moves to a state 260 wherein the first character in the first sequence is read and then to a state 262 wherein the first character of the second sequence is read. It should be understood that if the

sequence is a nucleotide sequence, then the character would normally be either A, T, C, G or U. If the sequence is a protein sequence, then it should be in the single letter amino acid code so that the first and sequence sequences can be easily compared.

A determination is then made at a decision state 264 whether the two characters are the same. If they are the same, then the process 250 moves to a state 268 wherein the next characters in the first and second sequences are read. A determination is then made whether the next characters are the same. If they are, then the process 250 continues this loop until two characters are not the same. If a determination is made that the next two characters are not the same, the process 250 moves to a decision state 274 to determine whether there are any more characters either sequence to read.

If there aren't any more characters to read, then the process 250 moves to a state 276 wherein the level of homology between the first and second sequences is displayed to the user. The level of homology is determined by calculating the proportion of characters between the sequences that were the same out of the total number of sequences in the first sequence. Thus, if every character in a first 100 nucleotide sequence aligned with a every character in a second sequence, the homology level would be 100%.

Alternatively, the computer program may be a computer program which compares the nucleotide sequences of the nucleic acid codes of the present invention, to reference nucleotide sequences in order to determine whether the nucleic acid code of the invention differs from a reference nucleic acid sequence at one or more positions. In some embodiments, such a program records the length and identity of inserted, deleted or substituted nucleotides with respect to the sequence of either the reference polynucleotide or the nucleic acid code of the invention. In one embodiment, the computer program may be a program which determines whether the nucleotide sequences of the nucleic acid codes of the invention contain one or more single nucleotide polymorphisms (SNP) with respect to a reference nucleotide sequence. These single nucleotide polymorphisms may each comprise a single base substitution, insertion, or deletion.

Another aspect of the present invention is a method for determining the level of homology between a polypeptide code of the invention and a reference polypeptide sequence, comprising the steps of reading the polypeptide code of the invention and the reference polypeptide sequence through use of a computer program which determines homology levels and determining homology between the polypeptide code and the reference polypeptide sequence using the computer program.

Accordingly, another aspect of the present invention is a method for determining whether a nucleic acid code of the invention differs at one or more nucleotides from a reference nucleotide sequence comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through use of a computer program which identifies differences between nucleic acid sequences and identifying differences between the nucleic acid code and the reference nucleotide sequence with the computer program. In some embodiments, the computer program is a program which identifies single

nucleotide polymorphisms The method may be implemented by the computer systems described above and the method illustrated in Figure 9. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of the invention and the reference nucleotide sequences through the use of the computer program and identifying differences between the nucleic acid codes and the reference nucleotide sequences with the computer program.

In other embodiments the computer based system may further comprise an identifier for identifying features within the nucleotide sequences of the nucleic acid codes of the invention or the amino acid sequences of the polypeptide codes of the invention.

An "identifier" refers to one or more programs which identifies certain features within the above-described nucleotide sequences of the nucleic acid codes of the invention or the amino acid sequences of the polypeptide codes of the invention. In one embodiment, the identifier may comprise a program which identifies an open reading frame in the cDNAs codes of the invention.

Figure 10 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence. The process 300 begins at a start state 302 and then moves to a state 304 wherein a first sequence that is to be checked for features is stored to a memory 115 in the computer system 100. The process 300 then moves to a state 306 wherein a database of sequence features is opened. Such a database would include a list of each feature's attributes along with the name of the feature. For example, a feature name could be "Initiation Codon" and the attribute would be "ATG". Another example would be the feature name "TAATAA Box" and the feature attribute would be "TAATAA". An example of such a database is produced by the University of Wisconsin Genetics Computer Group (www.gcg.com).

Once the database of features is opened at the state 306, the process 300 moves to a state 308 wherein the first feature is read from the database. A comparison of the attribute of the first feature with the first sequence is then made at a state 310. A determination is then made at a decision state 316 whether the attribute of the feature was found in the first sequence. If the attribute was found, then the process 300 moves to a state 318 wherein the name of the found feature is displayed to the user.

The process 300 then moves to a decision state 320 wherein a determination is made whether move features exist in the database. If no more features do exist, then the process 300 terminates at an end state 324. However, if more features do exist in the database, then the process 300 reads the next sequence feature at a state 326 and loops back to the state 310 wherein the attribute of the next feature is compared against the first sequence.

It should be noted, that if the feature attribute is not found in the first sequence at the decision state 316, the process 300 moves directly to the decision state 320 in order to determine if any more features exist in the database.

In another embodiment, the identifier may comprise a molecular modeling program which determines the 3-dimensional structure of the polypeptides codes of the invention. In some

embodiments, the molecular modeling program identifies target sequences that are most compatible with profiles representing the structural environments of the residues in known three-dimensional protein structures. (See, e.g., Eisenberg et al., U.S. Patent No 5,436,850 issued July 25, 1995). In another technique, the known three-dimensional structures of proteins in a given family are superimposed to define the structurally conserved regions in that family. This protein modeling technique also uses the known three-dimensional structure of a homologous protein to approximate the structure of the polypeptide codes of the invention. (See e.g., Srinivasan, et al., U.S. Patent No 5,557,535 issued September 17, 1996). Conventional homology modeling techniques have been used routinely to build models of proteases and antibodies. (Sowdhamini et al., Protein Engineering 10:207, 215 (1997)). Comparative approaches can also be used to develop three-dimensional protein models when the protein of interest has poor sequence identity to template proteins. In some cases, proteins fold into similar three-dimensional structures despite having very weak sequence identities. For example, the three-dimensional structures of a number of helical cytokines fold in similar three-dimensional topology in spite of weak sequence homology.

The recent development of threading methods now enables the identification of likely folding patterns in a number of situations where the structural relatedness between target and template(s) is not detectable at the sequence level. Hybrid methods, in which fold recognition is performed using Multiple Sequence Threading (MST), structural equivalencies are deduced from the threading output using a distance geometry program DRAGON to construct a low resolution model, and a full-atom representation is constructed using a molecular modeling package such as QUANTA.

According to this 3-step approach, candidate templates are first identified by using the novel fold recognition algorithm MST, which is capable of performing simultaneous threading of multiple aligned sequences onto one or more 3-D structures. In a second step, the structural equivalencies obtained from the MST output are converted into interresidue distance restraints and fed into the distance geometry program DRAGON, together with auxiliary information obtained from secondary structure predictions. The program combines the restraints in an unbiased manner and rapidly generates a large number of low resolution model confirmations. In a third step, these low resolution model confirmations are converted into full-atom models and subjected to energy minimization using the molecular modeling package QUANTA. (See e.g., Aszódi et al., Proteins:Structure, Function, and Genetics, Supplement 1:38-42 (1997)).

The results of the molecular modeling analysis may then be used in rational drug design techniques to identify agents which modulate the activity of the polypeptide codes of the invention.

Accordingly, another aspect of the present invention is a method of identifying a feature within the nucleic acid codes of the invention or the polypeptide codes of the invention comprising reading the nucleic acid code(s) or the polypeptide code(s) through the use of a computer program which identifies features therein and identifying features within the nucleic acid code(s) or

polypeptide code(s) with the computer program. In one embodiment, computer program comprises a computer program which identifies open reading frames. In a further embodiment, the computer program identifies structural motifs in a polypeptide sequence. In another embodiment, the computer program comprises a molecular modeling program. The method may be performed by reading a single sequence or at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of the invention or the polypeptide codes of the invention through the use of the computer program and identifying features within the nucleic acid codes or polypeptide codes with the computer program.

The nucleic acid codes of the invention or the polypeptide codes of the invention may be stored and manipulated in a variety of data processor programs in a variety of formats. For example, they may be stored as text in a word processing file, such as MicrosoftWORD or WORDPERFECT or as an ASCII file in a variety of database programs familiar to those of skill in the art, such as DB2, SYBASE, or ORACLE. In addition, many computer programs and databases may be used as sequence comparers, identifiers, or sources of reference nucleotide or polypeptide sequences to be compared to the nucleic acid codes of the invention or the polypeptide codes of the invention. The following list is intended not to limit the invention but to provide guidance to programs and databases which are useful with the nucleic acid codes of the invention or the polypeptide codes of the invention. The programs and databases which may be used include, but are not limited to: MacPattern (EMBL), DiscoveryBase (Molecular Applications Group), GeneMine (Molecular Applications Group), Look (Molecular Applications Group), MacLook (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul et al, 1990), FASTA (Pearson and Lipman, 1988), FASTDB (Brutlag et al., 1990), Catalyst (Molecular Simulations Inc.), Catalyst/SHAPE (Molecular Simulations Inc.), Cerius$^2$.DBAccess (Molecular Simulations Inc.), HypoGen (Molecular Simulations Inc.), Insight II, (Molecular Simulations Inc.), Discover (Molecular Simulations Inc.), CHARMm (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), DelPhi, (Molecular Simulations Inc.), QuanteMM, (Molecular Simulations Inc.), Homology (Molecular Simulations Inc.), Modeler (Molecular Simulations Inc.), ISIS (Molecular Simulations Inc.), Quanta/Protein Design (Molecular Simulations Inc.), WebLab (Molecular Simulations Inc.), WebLab Diversity Explorer (Molecular Simulations Inc.), Gene Explorer (Molecular Simulations Inc.), SeqFold (Molecular Simulations Inc.), the EMBL/Swissprotein database, the MDL Available Chemicals Directory database, the MDL Drug Data Report data base, the Comprehensive Medicinal Chemistry database, Derwents's World Drug Index database, the BioByteMasterFile database, the Genbank database, and the Genseqn database. Many other programs and data bases would be apparent to one of skill in the art given the present disclosure.

Motifs which may be detected using the above programs include sequences encoding leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences encoding signal peptides which direct the secretion of the encoded

proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

Throughout this application, various publications, patents and published patent applications are cited. The disclosures of these publications, patents and published patent specification referenced in this application are hereby incorporated by reference into the present disclosure to more fully describe the sate of the art to which this invention pertains.

## EXAMPLES

### Example 1

### Identification Of Biallelic Markers - DNA Extraction

Blood donors were from French Caucasian origin. They presented a sufficient diversity for being representative of a French heterogeneous population. The DNA from 100 unrelated and healthy individuals was extracted, pooled and tested for the detection of biallelic markers. The pool was constituted by mixing equivalent quantities of DNA from each individual.

30 ml of peripheral venous blood were taken from each donor in the presence of EDTA. Cells (pellet) were collected after centrifugation for 10 minutes at 2000 rpm. Red cells were lysed by a lysis solution (50 ml final volume : 10 mM Tris pH7.6; 5 mM $MgCl_2$; 10 mM NaCl). The solution was centrifuged (10 minutes, 2000 rpm) as many times as necessary to eliminate the residual red cells present in the supernatant, after resuspension of the pellet in the lysis solution.

The pellet of white cells was lysed overnight at 42°C with 3.7 ml of lysis solution composed of:

- 3 ml TE 10-2 (Tris-HCl 10 mM, EDTA 2 mM) / NaCl 0.4 M
- 200 µl SDS 10%
- 500 µl K-proteinase (2 mg K-proteinase in TE 10-2 / NaCl 0.4 M).

For the extraction of proteins, 1 ml saturated NaCl (6M) (1/3.5 v/v) was added. After vigorous agitation, the solution was centrifuged for 20 minutes at 10000 rpm.

For the precipitation of DNA, 2 to 3 volumes of 100% ethanol were added to the previous supernatant, and the solution was centrifuged for 30 minutes at 2000 rpm. The DNA solution was rinsed three times with 70% ethanol to eliminate salts, and centrifuged for 20 minutes at 2000 rpm. The pellet was dried at 37°C, and resuspended in 1 ml TE 10-1 or 1 ml water. The DNA concentration was evaluated by measuring the OD at 260 nm (1 unit OD = 50 µg/ml DNA).

To determine the presence of proteins in the DNA solution, the OD 260 / OD 280 ratio was determined. Only DNA preparations having a OD 260 / OD 280 ratio between 1.8 and 2 were used in the subsequent examples described below.

**Example 2**

**Identification Of Biallelic Markers: Amplification Of Genomic DNA By PCR**

The amplification of specific genomic sequences of the DNA samples of example 1 was carried out on the pool of DNA obtained previously. In addition, 10 individual samples were similarly amplified.

PCR assays were performed using the following protocol:

| | |
|---|---|
| Final volume | 25 µl |
| DNA | 2 ng/µl |
| $MgCl_2$ | 2 mM |
| dNTP (each) | 200 µM |
| primer (each) | 2.9 ng/µl |
| Ampli Taq Gold DNA polymerase | 0.05 unit/µl |
| PCR buffer (10x = 0.1 M TrisHCl pH8.3 0.5M KCl) | 1x |

Each pair of first primers is designed using the sequence information of the *BAP28* gene disclosed herein and the OSP software (Hillier & Green, 1991). This first pair of primers were about 20 nucleotides in length.

**Table 1**

| Amplicon | Position range of the amplicon in SEQ ID No 1 | | Primer name | Position range of amplification primer in SEQ ID No 1 | | Primer name | Complementary position range of amplification primer in SEQ ID No 1 | |
|---|---|---|---|---|---|---|---|---|
| 5-381 | 4840 | 5266 | B1 | 4840 | 4859 | C1 | 5249 | 5266 |
| 5-382 | 5307 | 5729 | B2 | 5307 | 5324 | C2 | 5710 | 5729 |
| 99-7190 | 12946 | 13488 | B3 | 12946 | 12963 | C3 | 13471 | 13488 |
| 99-7203 | 23482 | 23929 | B4 | 23482 | 23501 | C4 | 23909 | 23929 |
| 5-383 | 27887 | 28315 | B5 | 27887 | 27904 | C5 | 28296 | 28315 |
| 99-7205 | 29833 | 30288 | B6 | 29833 | 29853 | C6 | 30270 | 30288 |
| 5-384 | 32439 | 32877 | B7 | 32439 | 32457 | C7 | 32858 | 32877 |
| 5-379 | 48110 | 48460 | B8 | 48110 | 48127 | C8 | 48441 | 48460 |
| 5-380 | 49558 | 49977 | B9 | 49558 | 49577 | C9 | 49958 | 49977 |
| 5-366 | 50162 | 50583 | B10 | 50162 | 50180 | C10 | 50564 | 50583 |
| 5-370 | 50937 | 51359 | B11 | 50937 | 50955 | C11 | 51341 | 51359 |
| 5-373 | 53437 | 53858 | B12 | 53437 | 53455 | C12 | 53840 | 53858 |
| 5-375 | 53974 | 54394 | B13 | 53974 | 53993 | C13 | 54375 | 54394 |
| 5-376 | 54602 | 55021 | B14 | 54602 | 54619 | C14 | 55002 | 55021 |
| 5-377 | 55608 | 56043 | B15 | 55608 | 55625 | C15 | 56025 | 56043 |
| 5-14 | 59673 | 60100 | B16 | 59673 | 59692 | C16 | 60083 | 60100 |
| 5-11 | 60718 | 61137 | B17 | 60718 | 60737 | C17 | 61119 | 61137 |
| 5-202 | 66177 | 66608 | B23 | 66177 | 66194 | C23 | 66589 | 66608 |
| 99-1605 | 71723 | 72170 | B21 | 71723 | 71743 | C21 | 72150 | 72170 |
| 5-2 | 71735 | 72169 | B22 | 71735 | 71754 | C22 | 72150 | 72169 |
| 5-171 | 85485 | 85905 | B20 | 85485 | 85502 | C20 | 85887 | 85905 |
| 5-169 | 86184 | 86600 | B19 | 86184 | 86203 | C19 | 86581 | 86600 |
| 99-1572 | 86932 | 87574 | B18 | 86932 | 86952 | C18 | 87556 | 87574 |
| 5-403 | 91068 | 91417 | B24 | 91068 | 91085 | C24 | 91398 | 91417 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **in SEQ ID No 29** | | | | | | | | |
| 99-13790 | 1 | 454 | B25 | 1 | 20 | C25 | 434 | 454 |
| **in SEQ ID No 25** | | | | | | | | |
| 99-13798 | 1 | 447 | B26 | 1 | 20 | C26 | 427 | 447 |
| **in SEQ ID No 27** | | | | | | | | |
| 99-13808 | 1 | 546 | B27 | 1 | 20 | C27 | 526 | 546 |
| **in SEQ ID No 30** | | | | | | | | |
| 99-13809 | 1 | 444 | B28 | 1 | 21 | C28 | 424 | 444 |
| **in SEQ ID No 28** | | | | | | | | |
| 99-13810 | 1 | 476 | B29 | 1 | 18 | C29 | 458 | 476 |
| **in SEQ ID No 23** | | | | | | | | |
| 99-1585 | 1 | 546 | B30 | 1 | 20 | C30 | 527 | 546 |
| **in SEQ ID No 24** | | | | | | | | |
| 99-1587 | 1 | 396 | B31 | 1 | 21 | C31 | 377 | 396 |
| **in SEQ ID No 31** | | | | | | | | |
| 99-1597 | 1 | 693 | B32 | 1 | 19 | C32 | 675 | 693 |
| **in SEQ ID No 26** | | | | | | | | |
| 99-1601 | 1 | 506 | B33 | 1 | 18 | C33 | 486 | 506 |
| **in SEQ ID No 18** | | | | | | | | |
| 99-7177 | 1 | 504 | B34 | 1 | 20 | C34 | 484 | 504 |
| **in SEQ ID No 22** | | | | | | | | |
| 99-7182 | 1 | 531 | B35 | 1 | 20 | C35 | 511 | 531 |
| **in SEQ ID No 21** | | | | | | | | |
| 99-7186 | 1 | 528 | B36 | 1 | 19 | C36 | 510 | 528 |
| **in SEQ ID No 20** | | | | | | | | |
| 99-7193 | 1 | 542 | B37 | 1 | 20 | C37 | 522 | 542 |
| **in SEQ ID No 19** | | | | | | | | |
| 99-7212 | 1 | 492 | B38 | 1 | 20 | C38 | 472 | 492 |

Preferably, the primers contained a common oligonucleotide tail upstream of the specific bases targeted for amplification which was useful for sequencing.

Primers PU contain the following additional PU 5' sequence:

TGTAAAACGACGGCCAGT; primers RP contain the following RP 5' sequence: CAGGAAACAGCTATGACC. The primer containing the additional PU 5' sequence is listed in SEQ ID No 11. The primer containing the additional RP 5' sequence is listed in SEQ ID No 12.

The synthesis of these primers was performed following the phosphoramidite method, on a GENSET UFPS 24.1 synthesizer.

DNA amplification was performed on a Genius II thermocycler. After heating at 95°C for 10 min, 40 cycles were performed. Each cycle comprised: 30 sec at 95°C, 54°C for 1 min, and 30 sec at 72°C. For final elongation, 10 min at 72°C ended the amplification. The quantities of the amplification products obtained were determined on 96-well microtiter plates, using a fluorometer and Picogreen as intercalant agent (Molecular Probes).

**Example 3**

**Identification Of Biallelic Markers - Sequencing Of Amplified Genomic DNA And Identification Of Polymorphisms**

The sequencing of the amplified DNA obtained in example 2 was carried out on ABI 377 sequencers. The sequences of the amplification products were determined using automated dideoxy terminator sequencing reactions with a dye terminator cycle sequencing protocol. The products of the sequencing reactions were run on sequencing gels and the sequences were determined using gel image analysis (ABI Prism DNA Sequencing Analysis software (2.1.2 version)).

The sequence data were further evaluated to detect the presence of biallelic markers within the amplified fragments. The polymorphism search was based on the presence of superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position as described previously.

The localization of the biallelic markers on SEQ ID Nos 1, and 18 to 31 are as shown above in Table 2.

Also encompassed by the present invention are BAP28-related biallelic markers A1 to A58 described below in Table 2.

**Table 2**

| Amplicon | BM | Marker Name | Localization in BAP28 gene | Polymorphism | | BM position in SEQ ID No 1 | BM position in SEQ ID Nos 2, 3 & 4 |
|---|---|---|---|---|---|---|---|
| | | | | all1 | all2 | | |
| 5-381 | A1 | 5-381-133 | 5` regulatory region | A | G | 4972 | |
| 5-382 | A2 | 5-382-162 | Exon 2 | C | T | 5468 | 178 |
| 5-382 | A3 | 5-382-310 | Intron 2-3 | C | T | 5616 | |
| 5-382 | A4 | 5-382-316 | Intron 2-3 | G | C | 5622 | |
| 99-7190 | A5 | 99-7190-213 | Intron 6-7 | C | T | 13158 | |
| 99-7203 | A6 | 99-7203-282 | Intron 16-17 | A | T | 23761 | |
| 99-7203 | A7 | 99-7203-286 | Intron 16-17 | C | T | 23765 | |
| 5-383 | A8 | 5-383-42 | Intron 19-20 | A | G | 27928 | |
| 5-383 | A9 | 5-383-184 | Exon 20 | G | T | 28070 | 2677 |
| 99-7205 | A10 | 99-7205-228 | Intron 20-21 | A | G | 30061 | |
| 5-384 | A11 | 5-384-312 | Intron 21-22 | G | C | 32750 | |
| 5-379 | A12 | 5-379-80 | Intron 32-33 | A | C | 48189 | |
| 5-380 | A13 | 5-380-58 | Intron 33-34 | G | T | 49615 | |
| 5-380 | A14 | 5-380-59 | Intron 33-34 | C | T | 49616 | |
| 5-366 | A15 | 5-366-143 | Intron 34-35 | A | G | 50304 | |
| 5-370 | A16 | 5-370-197 | Exon 36 | A | G | 51133 | 5193 |
| 5-370 | A17 | 5-370-247 | Exon 36 | C | T | 51183 | 5243 |
| 5-373 | A18 | 5-373-98 | Intron 38-39 | C | T | 53534 | |
| 5-373 | A19 | 5-373-164 | Exon 39 | C | T | 53600 | 5673 |
| 5-373 | A20 | 5-373-222 | Exon 39 | A | G | 53658 | 5731 |
| 5-375 | A21 | 5-375-200 | Exon 41 | A | G | 54173 | 6011 |
| 5-375 | A22 | 5-375-259 | Intron 41-42 | C | T | 54232 | |
| 5-375 | A23 | 5-375-296 | Intron 41-42 | G | C | 54269 | |
| 5-375 | A24 | 5-375-399 | Intron 41-42 | G | C | 54372 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5-376 | A25 | 5-376-266 | Exon 42 | A | G | 54867 | 6162 |
| 5-377 | A26 | 5-377-82 | Intron 42-43 | C | T | 55689 | |
| 5-377 | A27 | 5-377-227 | Exon 43 | A | G | 55834 | 6271 |
| 5-14 | A28 | 5-14-165 | Intron 45-B' | A | G | 59937 | |
| 5-11 | A29 | 5-11-158 | Intron 45-B' | C | T | 60980 | |
| 5-202 | A36 | 5-202-117 | Intron 45-B' | A | T | 66492 | |
| 5-202 | A35 | 5-202-95 | Intron 45-B' | A | C | 66514 | |
| 99-1605 | A33 | 99-1605-112 | Intron 45-B' | A | G | 71834 | |
| 5-2 | A34 | 5-2-178 | Intron 45-B' | A | G | 71993 | |
| 5-171 | A32 | 5-171-204 | Intron 45-B' | A | G | 85702 | |
| 5-169 | A31 | 5-169-97 | Intron B'-A' | G | C | 86504 | |
| 99-1572 | A30 | 99-1572-440 | Intron B'-A' | A | G | 87135 | |
| 5-403 | A37 | 5-403-325 | Intron B'-A' | C | T | 91093 | |
| 5-403 | A38 | 5-403-294 | Intron B'-A' | A | G | 91124 | |
| 5-403 | A39 | 5-403-209 | Intron B'-A' | C | T | 91209 | |
| 5-403 | A40 | 5-403-156 | Exon A' | C | T | 91262 | 7935 in SEQ ID No 3<br>256 in SEQ ID No 4 |

| Amplicon | BM | Marker Name | Polymorphism | | BM position |
|---|---|---|---|---|---|
| | | | all1 | all2 | |
| 99-13790 | A41 | 99-13790-129 | C | T | 127 in SEQ ID No 29 |
| 99-13798 | A42 | 99-13798-284 | A | G | 283 in SEQ ID No 25 |
| 99-13808 | A43 | 99-13808-80 | A | T | 79 in SEQ ID No 27 |
| 99-13808 | A44 | 99-13808-268 | A | C | 266 in SEQ ID No 27 |
| 99-13808 | A45 | 99-13808-425 | G | C | 419 in SEQ ID No 27 |
| 99-13808 | A46 | 99-13808-455 | A | G | 453 in SEQ ID No 27 |
| 99-13809 | A47 | 99-13809-153 | A | G | 153 in SEQ ID No 30 |
| 99-13810 | A48 | 99-13810-214 | C | T | 212 in SEQ ID No 28 |
| 99-13810 | A49 | 99-13810-170 | A | T | 168 in SEQ ID No 28 |
| 99-1585 | A50 | 99-1585-373 | C | T | 372 in SEQ ID No 23 |
| 99-1587 | A51 | 99-1587-281 | A | G | 278 in SEQ ID No 24 |
| 99-1597 | A52 | 99-1597-162 | A | G | 162 in SEQ ID No 31 |
| 99-1601 | A53 | 99-1601-402 | A | T | 402 in SEQ ID No 26 |
| 99-7177 | A54 | 99-7177-81 | C | T | 81 in SEQ ID No 18 |
| 99-7182 | A55 | 99-7182-49 | C | T | 49 in SEQ ID No 22 |
| 99-7186 | A56 | 99-7186-212 | A | G | 212 in SEQ ID No 21 |
| 99-7193 | A57 | 99-7193-228 | G | C | 226 in SEQ ID No 20 |
| 99-7212 | A58 | 99-7212-346 | C | T | 345 in SEQ ID No 19 |

BM refers to "biallelic marker". All1 and all2 refer respectively to allele 1 and allele 2 of the biallelic marker.

The biallelic markers A16, A19, A21 and A25 are located in exonic sequence and give amino acid polymorphisms. Indeed, the codon comprising the marker A16 encodes either a serine or an asparagine in position 1694 of the SEQ ID No 5 ; the codon comprising the marker A19 encodes either an alanine or a valine in position 1854 of the SEQ ID No 5 ; the codon comprising the marker A21 encodes either an aspartic acid or an asparagine in position 1967 of the SEQ ID No 5 ; the

codon comprising the marker A25 encodes either a glycine or a glutamic acid in position 2017 of the SEQ ID No 5.

The Table 3 discloses the probes specific of each biallelic markers.

**Table 3**

| BM | Marker Name | Position range of probes in SEQ ID No 1 | | Probes |
|---|---|---|---|---|
| A1 | 5-381-133 | 4960 | 4984 | P1 |
| A2 | 5-382-162 | 5456 | 5480 | P2 |
| A3 | 5-382-310 | 5604 | 5628 | P3 |
| A4 | 5-382-316 | 5610 | 5634 | P4 |
| A5 | 99-7190-213 | 13146 | 13170 | P5 |
| A6 | 99-7203-282 | 23749 | 23773 | P6 |
| A7 | 99-7203-286 | 23753 | 23777 | P7 |
| A8 | 5-383-42 | 27916 | 27940 | P8 |
| A9 | 5-383-184 | 28058 | 28082 | P9 |
| A10 | 99-7205-228 | 30049 | 30073 | P10 |
| A11 | 5-384-312 | 32738 | 32762 | P11 |
| A12 | 5-379-80 | 48177 | 48201 | P12 |
| A13 | 5-380-58 | 49603 | 49627 | P13 |
| A14 | 5-380-59 | 49604 | 49628 | P14 |
| A15 | 5-366-143 | 50292 | 50316 | P15 |
| A16 | 5-370-197 | 51121 | 51145 | P16 |
| A17 | 5-370-247 | 51171 | 51195 | P17 |
| A18 | 5-373-98 | 53522 | 53546 | P18 |
| A19 | 5-373-164 | 53588 | 53612 | P19 |
| A20 | 5-373-222 | 53646 | 53670 | P20 |
| A21 | 5-375-200 | 54161 | 54185 | P21 |
| A22 | 5-375-259 | 54220 | 54244 | P22 |
| A23 | 5-375-296 | 54257 | 54281 | P23 |
| A24 | 5-375-399 | 54360 | 54384 | P24 |
| A25 | 5-376-266 | 54855 | 54879 | P25 |
| A26 | 5-377-82 | 55677 | 55701 | P26 |
| A27 | 5-377-227 | 55822 | 55846 | P27 |
| A28 | 5-14-165 | 59925 | 59949 | P28 |
| A29 | 5-11-158 | 60968 | 60992 | P29 |
| A36 | 5-202-117 | 66480 | 66504 | P36 |
| A35 | 5-202-95 | 66502 | 66526 | P35 |
| A33 | 99-1605-112 | 71822 | 71846 | P33 |
| A34 | 5-2-178 | 71981 | 72005 | P34 |
| A32 | 5-171-204 | 85690 | 85714 | P32 |
| A31 | 5-169-97 | 86492 | 86516 | P31 |
| A30 | 99-1572-440 | 87123 | 87147 | P30 |
| A37 | 5-403-325 | 91081 | 91105 | P37 |
| A38 | 5-403-294 | 91112 | 91136 | P38 |
| A39 | 5-403-209 | 91197 | 91221 | P39 |
| A40 | 5-403-156 | 91250 | 91274 | P40 |
| **BM** | **Marker Name** | **Position range of probes** | | **Probes** |
| A41 | 99-13790-129 | 115-139 in SEQ ID No 29 | | P41 |
| A42 | 99-13798-284 | 271-295 in SEQ ID No 25 | | P42 |
| A43 | 99-13808-80 | 67-91 in SEQ ID No 27 | | P43 |

| | | | |
|---|---|---|---|
| A44 | 99-13808-268 | 254-278 in SEQ ID No 27 | P44 |
| A45 | 99-13808-425 | 407-431 in SEQ ID No 27 | P45 |
| A46 | 99-13808-455 | 441-465 in SEQ ID No 27 | P46 |
| A47 | 99-13809-153 | 141-165 in SEQ ID No 30 | P47 |
| A48 | 99-13810-214 | 200-224 in SEQ ID No 28 | P48 |
| A49 | 99-13810-170 | 156-180 in SEQ ID No 28 | P49 |
| A50 | 99-1585-373 | 360-384 in SEQ ID No 23 | P50 |
| A51 | 99-1587-281 | 266-290 in SEQ ID No 24 | P51 |
| A52 | 99-1597-162 | 150-174 in SEQ ID No 31 | P52 |
| A53 | 99-1601-402 | 390-414 in SEQ ID No 26 | P53 |
| A54 | 99-7177-81 | 69-93 in SEQ ID No 18 | P54 |
| A55 | 99-7182-49 | 37-61 in SEQ ID No 22 | P55 |
| A56 | 99-7186-212 | 200-224 in SEQ ID No 21 | P56 |
| A57 | 99-7193-228 | 214-238 in SEQ ID No 20 | P57 |
| A58 | 99-7212-346 | 333-357 in SEQ ID No 19 | P58 |

**Example 4**

**Validation Of The Polymorphisms Through Microsequencing**

The biallelic markers identified in example 3 were further confirmed and their respective frequencies were determined through microsequencing. Microsequencing was carried out for each individual DNA sample described in Example 1.

Amplification from genomic DNA of individuals was performed by PCR as described above for the detection of the biallelic markers with the same set of PCR primers.

The preferred primers used in microsequencing were about 19 nucleotides in length and hybridized just upstream of the considered polymorphic base. According to the invention, the primers used in microsequencing are detailed in Table 4.

**Table 4**

| Marker Name | BM | Mis1 | Position range of microsequencing primer mis 1 in SEQ ID No 1 | | Mis2 | Complementary position range of microsequencing primer mis. 2 in SEQ ID No 1 | |
|---|---|---|---|---|---|---|---|
| 5-381-133 | A1 | D1 | 4953 | 4971 | E1 | 4973 | 4991 |
| 5-382-162 | A2 | D2 | 5449 | 5467 | E2 | 5469 | 5487 |
| 5-382-310 | A3 | D3 | 5597 | 5615 | E3 | 5617 | 5635 |
| 5-382-316 | A4 | D4 | 5603 | 5621 | E4 | 5623 | 5641 |
| 99-7190-213 | A5 | D5 | 13139 | 13157 | E5 | 13159 | 13177 |
| 99-7203-282 | A6 | D6 | 23742 | 23760 | E6 | 23762 | 23780 |
| 99-7203-286 | A7 | D7 | 23746 | 23764 | E7 | 23766 | 23784 |
| 5-383-42 | A8 | D8 | 27909 | 27927 | E8 | 27929 | 27947 |
| 5-383-184 | A9 | D9 | 28051 | 28069 | E9 | 28071 | 28089 |
| 99-7205-228 | A10 | D10 | 30042 | 30060 | E10 | 30062 | 30080 |
| 5-384-312 | A11 | D11 | 32731 | 32749 | E11 | 32751 | 32769 |
| 5-379-80 | A12 | D12 | 48170 | 48188 | E12 | 48190 | 48208 |
| 5-380-58 | A13 | D13 | 49596 | 49614 | E13 | 49616 | 49634 |
| 5-380-59 | A14 | D14 | 49597 | 49615 | E14 | 49617 | 49635 |
| 5-366-143 | A15 | D15 | 50285 | 50303 | E15 | 50305 | 50323 |
| 5-370-197 | A16 | D16 | 51114 | 51132 | E16 | 51134 | 51152 |
| 5-370-247 | A17 | D17 | 51164 | 51182 | E17 | 51184 | 51202 |
| 5-373-98 | A18 | D18 | 53515 | 53533 | E18 | 53535 | 53553 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5-373-164 | A19 | D19 | 53581 | 53599 | E19 | 53601 | 53619 |
| 5-373-222 | A20 | D20 | 53639 | 53657 | E20 | 53659 | 53677 |
| 5-375-200 | A21 | D21 | 54154 | 54172 | E21 | 54174 | 54192 |
| 5-375-259 | A22 | D22 | 54213 | 54231 | E22 | 54233 | 54251 |
| 5-375-296 | A23 | D23 | 54250 | 54268 | E23 | 54270 | 54288 |
| 5-375-399 | A24 | D24 | 54353 | 54371 | E24 | 54373 | 54391 |
| 5-376-266 | A25 | D25 | 54848 | 54866 | E25 | 54868 | 54886 |
| 5-377-82 | A26 | D26 | 55670 | 55688 | E26 | 55690 | 55708 |
| 5-377-227 | A27 | D27 | 55815 | 55833 | E27 | 55835 | 55853 |
| 5-14-165 | A28 | D28 | 59918 | 59936 | E28 | 59938 | 59956 |
| 5-11-158 | A29 | D29 | 60961 | 60979 | E29 | 60981 | 60999 |
| 5-202-117 | A36 | D36 | 66473 | 66491 | E36 | 66493 | 66511 |
| 5-202-95 | A35 | D35 | 66495 | 66513 | E35 | 66515 | 66533 |
| 99-1605-112 | A33 | D33 | 71815 | 71833 | E33 | 71835 | 71853 |
| 5-2-178 | A34 | D34 | 71974 | 71992 | E34 | 71994 | 72012 |
| 5-171-204 | A32 | D32 | 85683 | 85701 | E32 | 85703 | 85721 |
| 5-169-97 | A31 | D31 | 86485 | 86503 | E31 | 86505 | 86523 |
| 99-1572-440 | A30 | D30 | 87116 | 87134 | E30 | 87136 | 87154 |
| 5-403-325 | A37 | D37 | 91074 | 91092 | E37 | 91094 | 91112 |
| 5-403-294 | A38 | D38 | 91105 | 91123 | E38 | 91125 | 91143 |
| 5-403-209 | A39 | D39 | 91190 | 91208 | E39 | 91210 | 91228 |
| 5-403-156 | A40 | D40 | 91243 | 91261 | E40 | 91263 | 91281 |

| **Marker Name** | **BM** | **Mis1** | **Position range of microsequencing primer mis 1** | **Mis2** | **Complementary position range of microsequencing primer mis. 2** |
|---|---|---|---|---|---|
| 99-13790-129 | A41 | D41 | 108-126 in SEQ ID No 29 | E41 | 128-146 in SEQ ID No 29 |
| 99-13798-284 | A42 | D42 | 264-282 in SEQ ID No 25 | E42 | 284-302 in SEQ ID No 25 |
| 99-13808-80 | A43 | D43 | 60-78 in SEQ ID No 27 | E43 | 80-98 in SEQ ID No 27 |
| 99-13808-268 | A44 | D44 | 247-265 in SEQ ID No 27 | E44 | 267-285 in SEQ ID No 27 |
| 99-13808-425 | A45 | D45 | 400-418 in SEQ ID No 27 | E45 | 420-438 in SEQ ID No 27 |
| 99-13808-455 | A46 | D46 | 434-452 in SEQ ID No 27 | E46 | 454-472 in SEQ ID No 27 |
| 99-13809-153 | A47 | D47 | 134-152 in SEQ ID No 30 | E47 | 154-172 in SEQ ID No 30 |
| 99-13810-214 | A48 | D48 | 193-211 in SEQ ID No 28 | E48 | 213-231 in SEQ ID No 28 |
| 99-13810-170 | A49 | D49 | 149-167 in SEQ ID No 28 | E49 | 169-187 in SEQ ID No 28 |
| 99-1585-373 | A50 | D50 | 353-371 in SEQ ID No 23 | E50 | 373-391 in SEQ ID No 23 |
| 99-1587-281 | A51 | D51 | 259-277 in SEQ ID No 24 | E51 | 279-297 in SEQ ID No 24 |
| 99-1597-162 | A52 | D52 | 143-161 in SEQ ID No 31 | E52 | 163-181 in SEQ ID No 31 |
| 99-1601-402 | A53 | D53 | 383-401 in SEQ ID No 26 | E53 | 403-421 in SEQ ID No 26 |
| 99-7177-81 | A54 | D54 | 62-80 in SEQ ID No 18 | E54 | 82-100 in SEQ ID No 18 |
| 99-7182-49 | A55 | D55 | 30-48 in SEQ ID No 22 | E55 | 50-68 in SEQ ID No 22 |
| 99-7186-212 | A56 | D56 | 193-211 in SEQ ID No 21 | E56 | 213-231 in SEQ ID No 21 |
| 99-7193-228 | A57 | D57 | 207-225 in SEQ ID No 20 | E57 | 227-245 in SEQ ID No 20 |
| 99-7212-346 | A58 | D58 | 326-344 in SEQ ID No 19 | E58 | 346-364 in SEQ ID No 19 |

Mis 1 and Mis 2 respectively refer to microsequencing primers which hybridized with the non-coding strand of the *BAP28* gene or with the coding strand of the *BAP28* gene.

The microsequencing reaction was performed as follows :

After purification of the amplification products, the microsequencing reaction mixture was prepared by adding, in a 20µl final volume: 10 pmol microsequencing oligonucleotide, 1 U

Thermosequenase (Amersham E79000G), 1.25 µl Thermosequenase buffer (260 mM Tris HCl pH 9.5, 65 mM $MgCl_2$), and the two appropriate fluorescent ddNTPs (Perkin Elmer, Dye Terminator Set 401095) complementary to the nucleotides at the polymorphic site of each biallelic marker tested, following the manufacturer's recommendations. After 4 minutes at 94°C, 20 PCR cycles of 15 sec at 55°C, 5 sec at 72°C, and 10 sec at 94°C were carried out in a Tetrad PTC-225 thermocycler (MJ Research). The unincorporated dye terminators were then removed by ethanol precipitation. Samples were finally resuspended in formamide-EDTA loading buffer and heated for 2 min at 95°C before being loaded on a polyacrylamide sequencing gel. The data were collected by an ABI PRISM 377 DNA sequencer and processed using the GENESCAN software (Perkin Elmer).

Following gel analysis, data were automatically processed with software that allows the determination of the alleles of biallelic markers present in each amplified fragment.

The software evaluates such factors as whether the intensities of the signals resulting from the above microsequencing procedures are weak, normal, or saturated, or whether the signals are ambiguous. In addition, the software identifies significant peaks (according to shape and height criteria). Among the significant peaks, peaks corresponding to the targeted site are identified based on their position. When two significant peaks are detected for the same position, each sample is categorized classification as homozygous or heterozygous type based on the height ratio.

## Example 5

## Association Study Between Prostate Cancer And The Biallelic Markers Of The *PCTA-1* Gene

### Collection Of DNA Samples From Affected And Non-Affected Individuals

Affected population :

The positive trait followed in this association study was prostate cancer. Prostate cancer patients were recruited according to a combination of clinical, histological and biological inclusion criteria. Clinical criteria can include rectal examination and prostate biopsies. Biological criteria can include PSA assays. The affected individuals were recorded as familial forms when at least two persons affected by prostate cancer have been diagnosed in the family. Remaining cases were classified as sporadic cases, and more particularly in informative cases (at least two sibs of the case both aged over 50 years old are unaffected), or sporadic uninformative cases (no information about sibs over 50 years old is available). All affected individuals included in the statistical analysis of this patent were unrelated. Cases were also separated following the criteria of diagnosis age : early onset prostate cancer (under 65 years old) and late onset prostate cancer (65 years old or more).

Unaffected population :

Control individuals included in this study were checked for both the absence of all clinical and biological criteria defining the presence or the risk of prostate cancer (PSA < 4) (WO 96/21042), and for their age (aged 65 years old or more). All unaffected individuals included in the statistical analysis of this patent were unrelated.

The affected group was composed by 491 unrelated individuals, comprising:

- 197 familial cases; and
- 294 sporadic cases, 70 of which are sporadic informative cases.

The unaffected group contained 313 individuals which were 65 years or older.

**Genotyping Of Affected And Control Individuals**

The general strategy to perform the association studies was to individually scan the DNA samples from all individuals in each of the populations described above in order to establish the allele frequencies of the above described biallelic markers in each of these populations. More particularly, the 30 biallelic markers used in the present association study are described in Table 5.

Allelic frequencies of the biallelic markers of the Table 5 in each population were determined by performing microsequencing reactions on amplified fragments obtained by genomic PCR performed on the DNA samples from each individual. Genomic PCR and microsequencing were performed as detailed above in examples 2 and 4 using the described PCR and microsequencing primers.

**Table 5**

| BM | Marker Name | Position in *BAP28* gene | Position in *PCTA-1* gene | Nb of controls | Frequency (allele) |
|---|---|---|---|---|---|
| A54 | 99-7177/81 | 5' of gene | 3' of gene | 257 | 69.07 (C) |
| A58 | 99-7212/346 | 5' of gene | 3' of gene | 259 | 66.99 (C) |
| A57 | 99-7193/228 | 5' of gene | 3' of gene | 250 | 59.2 (C) |
| A56 | 99-7186/212 | 5' of gene | 3' of gene | 292 | 66.1 (A) |
| A55 | 99-7182/49 | 5' of gene | 3' of gene | 287 | 63.59 (C) |
| A1 | 5-381/133 | 5`regulatory region | 3' of gene | 304 | 65.46 (G) |
| A4 | 5-382/316 | intron 2-3 | 3' of gene | 304 | 65.79 (C) |
| A5 | 99-7190/213 | intron 6-7 | 3' of gene | 297 | 72.9 (C) |
| A7 | 99-7203/286 | intron 16-17 | 3' of gene | 257 | 68.09 (T) |
| A11 | 5-384/312 | intron 21-22 | 3' of gene | 211 | 73.22 (G) |
| A12 | 5-379/80 | intron 32-33 | 3' of gene | 294 | 73.98 (A) |
| A16 | 5-370/197 | Exon 36 | 3' of gene | 287 | 76.31 (G) |
| A19 | 5-373/164 | Exon 39 | 3' of gene | 298 | 68.62 (C) |
| A21 | 5-375/200 | exon 41 | 3' of gene | 307 | 68.73 (G) |
| A25 | 5-376/266 | exon 42 | 3' of gene | 298 | 68.96 (G) |
| A27 | 5-377/227 | exon 43 | 3' of gene | 307 | 68.73 (A) |
| A28 | 5-14/165 | intron 45-B` | 3' UTR | 307 | 65.15 (T) |
| A29 | 5-11/158 | intron 45-B` | 3' UTR | 303 | 75.41 (G) |
| A35 | 5-202/95 | intron 45-B` | Exon 6b | 308 | 95.13 (G) |
| A33 | 99-1605/112 | intron 45-B` | intron 2 | 304 | 68.75 (G) |
| A34 | 5-2/178 | intron 45-B` | Exon 2 | 306 | 68.3 (C) |
| A32 | 5-171/204 | intron 45-B` | intron B | 307 | 70.85 (T) |
| A31 | 5-169/97 | intron B`-A` | intron D | 305 | 82.3 (C) |
| A30 | 99-1572/440 | intron B`-A` | intron D | 304 | 65.79 (T) |
| A50 | 99-1585/373 | 3` of gene | 5' of gene | 300 | 78 (C) |
| A51 | 99-1587/281 | 3` of gene | 5' of gene | 286 | 67.31 (G) |
| A42 | 99-13798/284 | 3` of gene | 5' of gene | 278 | 53.42 (A) |
| A53 | 99-1601/402 | 3` of gene | 5' of gene | 305 | 67.21 (A) |
| A43 | 99-13808/80 | 3` of gene | 5' of gene | 214 | 59.58 (T) |

| | | | | | |
|---|---|---|---|---|---|
| A48 | 99-13810/214 | 3' of gene | 5' of gene | 289 | 59.86 (T) |

**Association Study Between Prostate Cancer And The Biallelic Markers Of The BAP28 Gene : Single marker association**

Frequencies of biallelic alleles were compared in case-control populations described above. We compare different sub-populations in function of phenotypes (sporadic and familial cases vs controls) to determine the characterisation of assocation.

The Figure 5 shows the results of allelic association analysis for markers localized in and around *BAP28* gene. This analysis tests the difference of allelic frequency for each marker between population. The statistical significance of this difference is assessed by performing a Pearson chi-square test with one degree of freedom.

The genotyped markers A55 (99-7182/49), A4 (5-382/316), A19 (5-373/164), A28 (5-14/165), A42 (99-13798/284), and A53 (99-1601/402) are significant at the 5% level for allelic test (respectively, pvalue=4 x $10^{-2}$, 4 x $10^{-3}$, 4 x $10^{-2}$, 1 x $10^{-2}$, 2 x $10^{-2}$, and 7 x $10^{-3}$) for sporadic cases. The 4 markers A28 (5-14/165), A4 (5-382/316), A1 (5-381/133), and A55 (99-7182/49) present a high significant association for allelic test (respectively, pvalue=4 x $10^{-5}$, 8 x $10^{-6}$, 3 x $10^{-5}$, and 1 x $10^{-4}$) between informatif sporadic cases and controls. The marker A30 (99-1572/440) is significant for familial cases (allelic pvalue=3 x $10^{-2}$).

Frequencies of the genotypes for one biallelic marker were compared in case-control populations described above. We compare different sub-populations in function of phenotypes (sporadic and familial cases vs controls) to determine the characterisation of assocation. The Figure 6 shows the results of genotypic association analysis for markers localized in and around *BAP28* gene. This analysis compares the three genotype frequencies between the two studied population. The statistical test used is a Pearson chi-square with 2 degree of freedom.

The genotyped markers A4 (5-382/316), A19 (5-373/164), A28 (5-14/165), A50 (99-1585/373), A42 (99-13798/284), and A53 (99-1601/402) are significant at the 5% level for allelic test (respectively, pvalue=9 x $10^{-3}$, 9 x $10^{-2}$, 4 x $10^{-2}$, 4 x $10^{-2}$, 8 x $10^{-2}$, and 3 x $10^{-2}$) for sporadic cases. The 4 markers A28 (5-14/165), A4 (5-382/316), A1 (5-381/133), and A55 (99-7182/49) present a high significant association for allelic test (respectively, pvalue=1 x $10^{-5}$, 2 x $10^{-5}$, 3 x $10^{-6}$, and 1 x $10^{-5}$) between informatif sporadic cases and controls. The 2 markers A31 (5-169/97) and A33 (99-1605/112) are significant for familial cases (respectively, pvalue=3 x $10^{-2}$ and 2 x $10^{-2}$).

The results of the association studies show that a polymorphism of the *BAP28* gene is related to sporadic and/or familial assocation. The biallelic markers A55 (99-7182/49), A1 (5-381/133), A4 (5-382/316), A19 (5-373/164), A28 (5-14/165), A50 (99-1585/373), A42 (99-13798/284), A31 (5-169/97), A33 (99-1605/112), and A53 (99-1601/402) can be then used in diagnostics with a test based on these markers.

**Haplotype Frequency Analysis**

One way of increasing the statistical power of individual markers, is by performing haplotype association analysis.

Haplotype analysis for association of *BAP28* markers and prostate cancer was performed by estimating the frequencies of all possible haplotypes comprising biallelic markers of the Table 5 in the cases and control populations described in Example 5, and comparing these frequencies by means of a chi square statistical test (one degree of freedom). Haplotype estimations were performed by applying the Expectation-Maximization (EM) algorithm (Excoffier L & Slatkin M, 1995), using the EM-HAPLO program (Hawley ME, Pakstis AJ & Kidd KK, 1994). More particularly, two tests were performed, namely a haplo-max test and an Omnibus LR test which compares the profile of haplotype frequencies were also performed.

The haplo-max test, which is based on haplotype frequencies differences, selects the difference showing the maximum positive (maxM) or negative (maxS) test value between cases versus controls (rejecting test values based on rare haplotype frequencies, i.e, with an estimated number of haplotypes carriers inferior to 10) ; for one combination of markers there is therefore one Max-M and one Max-S test values.

For one combination of 2, 3 or 4 markers, the Omnibus Likelihood ratio test allows to compare the profile of haplotype frequency differences between the two populations under study. The null hypothesis is that both cases and controls are samples derived from the same population, i.e., the haplotypes frequencies are close. Using the E-M algorithm, one can calculate the haplotype frequencies in cases, in controls and in the overall population. Once the haplotype frequencies are estimated, a likelihood ratio test (LR test) can be derived. It has to be underlined that for one combination of markers, only one LR test is obtained. If the data at hand would be observed haplotypes frequencies, provided there are no rare haplotypes, the LR test should follows a Chi-square distribution with h-1 degree of freedom, h being the number of possible haplotypes. This is to say: for two markers, a chi-square with 4 degree of freedom; for 3 markers, a chi-square with 7 degree of freedom; and for 4-markers, a chi-square with 15 degree of freedom. As haplotype frequencies are only inferred via the E-M algorithm and that rare haplotypes occur, a permutation procedure is more suitable.

The results of haplotype analysis using all combinations of 2 or 3 biallelic markers from the *BAP28*-related biallelic markers of the Table 5 are represented in the Figures 7 to 11. As above-mentioned, the profile of haplotypes frequencies have been compared by two main approaches: Individual haplotype tests and Omnibus Likelihood ratio tests. A permutation procedure allowed assessment of the significance of the tests. The most significant haplotypes obtained are shown in Figure 12. We analyzed separately the familial cases and sporadic cases, because the singlepoint analyses showed the different significant SNPs pattern.

**Haplotype frequency analysis for prostate cancer cases**

The most significant haplotypes obtained with the cases of prostate cancer are shown in Figure 7 a and b.

The two-markers haplotypes comprise the biallelic markers A1 (5-381/133), A4 (5-382/316), A19 (5-373/164), A21 (5-375/200), A25 (5-376/266), A27 (5-377/227), A53 (99-1601/402), A42 (99-13798/284), and A55 (99-7182/49).

The preferred two-markers haplotypes are described in Figure 7a as H1 to H8. All these haplotypes comprise either the biallelic marker A53 (99-1601/402) or A42 (99-13798/284). One of the more preferred haplotype is the haplotype H1 and it comprises the biallelic markers A53 (99-1601/402) and A27 (5-377/227), alleles TG respectively. This haplotype presented a p-value for the haplotype frequency test of $3.9 \times 10^{-4}$ and an odd-ratio of 1.80. Estimated haplotype frequencies were 15.6 % in the cases and 9.3 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $1.7 \times 10^{-2}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $5 \times 10^{-2}$ by omnibus Lr test.

The three-markers haplotypes comprise the biallelic markers A53 (99-1601/402), A42 (99-13798/284), A51 (99-1587/281), A31 (5-169/97), A34 (5-2/178), A33 (99-1605/112), A28 (5-14/165), A27 (5-377/227), A25 (5-376/266), A21 (5-375/200), A19 (5-373/164), A7 (99-7203/286), A4 (5-382/316), A55 (99-7182/49), A56 (99-7186/212), A57 (99-7193/228), A58 (99-7212/346).

The preferred three-markers haplotypes are described in Figure 7b as H435 to H452. All these haplotypes comprise the biallelic marker A53 (99-1601/402). Most of them comprise the biallelic marker A51 (99-1587/281). The more preferred haplotype is the haplotype H435 and comprises the biallelic markers A53 (99-1601/402), A51 (99-1587/281) and A34 (5-2/178), alleles TAT, respectively. This haplotype presented a p-value for the haplotype frequency test of $3.3 \times 10^{-8}$ and an odd-ratio of 100. Estimated haplotype frequencies were 5.3 % in the cases and 0 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $7.3 \times 10^{-3}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test.

In conclusion, most preferred haplotypes for the cases of prostate cancer comprise the biallelic marker A53 (99-1601/402). Some other preferred haplotypes for the cases of prostate cancer comprise the biallelic markers A42 (99-13798/284) and/or A51 (99-1587/281). These haplotypes can be used in diagnostic, more particularly in diagnostics of prostate cancer susceptibility.

**Haplotype frequency analysis for familial cases of prostate cancer**

The most significant haplotypes obtained with the familial cases of prostate cancer are shown in Figure 8 a and b.

The two-markers haplotypes comprise the biallelic markers A51 (99-1587/281), A30 (99-1572/440), A32 (5-171/204), A34 (5-2/178), A33 (99-1605/112), A29 (5-11/158), A27 (5-377/227), A19 (5-373/164), A5 (99-7190/213), A56 (99-7186/212), and A54 (99-7177/81).

The preferred two-markers haplotypes are described in Figure 8a as H1 to H10. All these haplotypes comprise either the biallelic marker A51 (99-1587/281) or A30 (99-1572/440). One of the more preferred haplotype is the haplotype H4. The pvalue of haplotype H 4 obtained by a chi-square distribution with 2 ddl for this combination of 2 markers with A30 (99-1572/440) and A32 (5-171/204) is $2.4 \times 10^{-3}$ by omnibus test. These markers are not in disequilibium linkage. In concerning the individual haplotype test, this haplotype consisting of 2 biallelic markers presented a $9.7 \times 10^{-5}$ p-value of and an odd-ratio of 1.7, for alleles TT respectively. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test. This haplotype tested on all cases-controls population gives estimated haplotype frequencies for sporadic cases (n=197) of 57.1% and for controls (n=313) of 44.1%. The trend about of estimations of haplotype frequencies are not identic between familial and sporadic cases , but the trend of sporadics are same for controls.

The three-markers haplotypes comprise the biallelic markers A48 (99-13810/214), A53 (99-1601/402), A42 (99-13798/284), A51 (99-1587/281), A30 (99-1572/440), A32 (5-171/204), A34 (5-2/178), A33 (99-1605/112), A29 (5-11/158), A27 (5-377/227), A19 (5-373/164), A7 (99-7203/286), A5 (99-7190/213), A56 (99-7186/212) and A54 (99-7177/81).

The preferred three-markers haplotypes are described in Figure 8b as H436 to H454. Most of them comprise the biallelic marker A30 (99-1572/440), A51 (99-1587/281) and A53 (99-1601/402). One of the more preferred haplotype is the haplotype H437 and comprises the biallelic markers A53 (99-1601/402), A30 (99-1572/440) and A54 (99-7177/81), alleles ATC, respectively. This haplotype presented a p-value for the haplotype frequency test of $3.6 \times 10^{-7}$ and an odd-ratio of 2.13. Estimated haplotype frequencies were 44.8 % in the cases and 27.6 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $2.9 \times 10^{-3}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test.

In conclusion, most preferred haplotypes for the familial cases of prostate cancer comprise the biallelic markers A30 (99-1572/440), and A51 (99-1587/281). These haplotypes can be used in diagnostic, more particularly in diagnostics of familial prostate cancer susceptibility.

The most significant haplotypes obtained with the early onset familial cases of prostate cancer are shown in Figure 9 a and b.

The two-markers haplotypes comprise the biallelic markers A42 (99-13798/284), A51 (99-1587/281), A50 (99-1585/373), A30 (99-1572/440), A32 (5-171/204), A34 (5-2/178), A33 (99-1605/112), A29 (5-11/158), A19 (5-373/164), A16 (5-370/197), A12 (5-379/80), A11 (5-384/312), A7 (99-7203/286), A5 (99-7190/213), A4 (5-382/316), and A54 (99-7177/81).

The preferred two-markers haplotypes are described in Figure 7a as H1 to H13. Most of these haplotypes comprise the biallelic marker A30 (99-1572/440). One of the more preferred haplotype is the haplotype H1 and it comprises the biallelic markers A30 (99-1572/440) and A32 (5-171/204), alleles TT respectively. This haplotype presented a p-value for the haplotype frequency test of $2.5 \times 10^{-6}$ and an odd-ratio of 2.28. Estimated haplotype frequencies were 64.4 % in the cases and 44.2 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $8.3 \times 10^{-5}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $5 \times 10^{-2}$ by omnibus Lr test.

The three-markers haplotypes comprise the biallelic markers A53 (99-1601/402), A30 (99-1572/440), A32 (5-171/204), A34 (5-2/178), A33 (99-1605/112), A29 (5-11/158), A21 (5-375/200), A19 (5-373/164), A12 (5-379/80), A11 (5-384/312), A7 (99-7203/286), A5 (99-7190/213), A56 (99-7186/212), and A54 (99-7177/81).

The preferred three-markers haplotypes are described in Figure 9b as H421 to H443. All of them comprise the biallelic marker A30 (99-1572/440)and almost all of them comprise the biallelic marker A53 (99-1601/402). One of the more preferred haplotype is the haplotype H421 and comprises the biallelic markers A53 (99-1601/402), A30 (99-1572/440) and A5 (99-7190/213), alleles ATC, respectively. This haplotype presented a p-value for the haplotype frequency test of $2.3 \times 10^{-7}$ and an odd-ratio of 2.7. Estimated haplotype frequencies were 52.3 % in the cases and 28.8 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $8.6 \times 10^{-4}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test.

In conclusion, most preferred haplotypes for the early onset familial cases of prostate cancer comprise the biallelic markers A30 (99-1572/440), and A53 (99-1601/402). These haplotypes can be used in diagnostic, more particularly in diagnostics of early onset familial prostate cancer susceptibility.

**Haplotype frequency analysis for sporadic cases of prostate cancer**

The most significant haplotypes obtained with the sporadic cases of prostate cancer are shown in Figure 10 a and b.

The two-markers haplotypes comprise the biallelic markers A53 (99-1601/402), A42 (99-13798/284), A32 (5-171/204), A29 (5-11/158), A28 (5-14/165), A27 (5-377/227), A25 (5-376/266), A19 (5-373/164), A16 (5-370/197), A4 (5-382/316), and A55 (99-7182/49).

The preferred two-markers haplotypes are described in Figure 10a as H1 to H12. The more usual biallelic markers in these haplotypes are A4 (5-382/316), A53 (99-1601/402), and A42 (99-13798/284). One of the more preferred haplotype is the haplotype H1 and comprises the biallelic markers A53 (99-1601/402), and A4 (5-382/316), alleles TG respectively. This haplotype presented a p-value for the haplotype frequency test of $1 \times 10^{-5}$ and an odd-ratio of 2.09. Estimated haplotype

frequencies were 19.9 % in the cases and 10.6 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $4.4 \times 10^{-4}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test. The results of allelic association which show that these markers are associated are significant. The haplotype analysis by combining the informativeness of a set of biallelic markers increases the power of the association analysis, allowing false positive and/or negative data that may result from the single marker studies to be elimated. . The significant trend for singlepoint analysis seems to be identic for multipoint analysis. This haplotype tested on all cases-controls population gives estimated haplotype frequencies for sporadic cases (n=294) of 19.6% and for controls (n=313) of 10.6%. For the same haplotype, any significant results for familial cases can be found. Therefore, the association for sporadic cases is differents for familial cases.

The three-markers haplotypes comprise the biallelic markers A53 (99-1601/402), A42 (99-13798/284), A51 (99-1587/281), A31 (5-169/97), A34 (5-2/178), A27 (5-377/227), A25 (5-376/266), A21 (5-375/200), A19 (5-373/164),and A55 (99-7182/49).

The preferred three-markers haplotypes are described in Figure 10b as H436 to H444. All the haplotypes comprise the biallelic marker A53 (99-1601/402). The biallelic markers A42 (99-13798/284) and A51 (99-1587/281) are frequently found in these haplotypes. One of the more preferred haplotype is the haplotype H436 and comprises the biallelic markers A53 (99-1601/402), A51 (99-1587/281) and A34 (5-2/178), alleles TAT respectively. This haplotype presented a p-value for the haplotype frequency test of $5.4 \times 10^{-7}$ and an odd-ratio of 100. Estimated haplotype frequencies were 5.6 % in the cases and 0 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $3.5 \times 10^{-3}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test..

In conclusion, most preferred haplotypes for the sporadic cases of prostate cancer comprise the biallelic marker A53 (99-1601/402). The biallelic markers A42 (99-13798/284), A51 (99-1587/281) and A4 (5-382/316) are frequently found in the preferred haplotypes. These haplotypes can be used in diagnostic, more particularly in diagnostics of sporadic prostate cancer susceptibility.

The most significant haplotypes obtained with the informative sporadic cases of prostate cancer are shown in Figure 11 a and b.

The two-markers haplotypes comprise the biallelic markers A53 (99-1601/402), A30 (99-1572/440), A32 (5-171/204), A29 (5-11/158), A16 (5-370/197), A4 (5-382/316), A1 (5-381/133), and A55 (99-7182/49).

The preferred two-markers haplotypes are described in Figure 11a as H1 to H11. The more usual biallelic markers in these haplotypes are A4 (5-382/316), and A1 (5-381/133). One of the more preferred haplotype is the haplotype H1 and comprises the biallelic markers A16 (5-370/197),

and A1 (5-381/133), alleles GA respectively. This haplotype presented a p-value for the haplotype frequency test of $9.4 \times 10^{-8}$ and an odd-ratio of 3.43. Estimated haplotype frequencies were 28.6 % in the cases and 10.5 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $6.7 \times 10^{-7}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test.

The three-markers haplotypes comprise the biallelic markers A53 (99-1601/402), A50 (99-1585/373), A30 (99-1572/440), A31 (5-169/97), A34 (5-2/178), A33 (99-1605/112), A29 (5-11/158), A28 (5-14/165), A27 (5-377/227), A25 (5-376/266), A21 (5-375/200), A16 (5-370/197), A4 (5-382/316), A1 (5-381/133), and A55 (99-7182/49).

The preferred three-markers haplotypes are described in Figure 11b as H415 to H430. Most of the haplotypes comprise the biallelic markers A53 (99-1601/402) and A31 (5-169/97). The biallelic markers A50 (99-1585/373), A16 (5-370/197), A4 (5-382/316), and A1 (5-381/133) are frequently found in these haplotypes. One of the more preferred haplotype is the haplotype H415 and comprises the biallelic markers A50 (99-1585/373), A16 (5-370/197), and A1 (5-381/133), alleles CGA respectively. This haplotype presented a p-value for the haplotype frequency test of $3.8 \times 10^{-9}$ and an odd-ratio of 4.25. Estimated haplotype frequencies were 26.7 % in the cases and 7.9 % in the controls. This haplotype presented a p-value for the likelihood ratio test of $3.3 \times 10^{-6}$. The pvalue by permutation test is $<1 \times 10^{-2}$ and the pvalue for this group of markers is $1 \times 10^{-2}$ by omnibus Lr test..

In conclusion, most preferred haplotypes for the informative sporadic cases of prostate cancer comprise the biallelic markers A53 (99-1601/402), A31 (5-169/97), A4 (5-382/316), and A1 (5-381/133). The biallelic markers A50 (99-1585/373), A16 (5-370/197) are also frequently found in the preferred haplotypes. These haplotypes can be used in diagnostic, more particularly in diagnostics of informative sporadic prostate cancer susceptibility.

**Summary of haplotype frequency analysis**

The most preferred two- biallelic markers haplotypes for the familial and sporadic prostate cancer are summarized in Figure 12. This haplotype can be used in diagnostic of prostate cancer susceptibility.

The statistical significance of the results obtained for the haplotype analysis was evaluated by a phenotypic permutation test reiterated 1000 times on a computer. For this computer simulation, data from the cases and control individuals were pooled and randomly allocated to two groups which contained the same number of individuals as the case-control populations used to produce the haplotype frequency analysis data. A haplotype analysis was then run on these artificial groups for the preferred haplotypes which presented a strong association with prostate cancer. This experiment was reiterated 1000 times and the results are shown in Figure 12.

Figure 12A shows the association results the preferred haplotype with A30 (99-1572/440) and A32 (5-171/204), alleles TT, for each population and with 1000 permutations. This haplotype is specific of familial prostate cancer, and more particularly of early onset prostate cancer. This haplotype is highly significant and could be used in diagnostic.

Figure 12B shows the association results the preferred haplotype with A16 (5-370/197), and A1 (5-381/133), alleles GA, for each population and with 1000 permutations. This haplotype is specific of sporadic prostate cancer. This haplotype is highly significant and could be used in diagnostic.

Figure 12C shows the association results the preferred haplotype with A53 (99-1601/402), and A4 (5-382/316), alleles TG, for each population and with 1000 permutations. This haplotype is specific of prostate cancer, and more particularly of sporadic prostate cancer. This haplotype is highly significant and could be used in diagnostic.

**Example 6**

**Preparation of Antibody Compositions to the BAP28 protein**

Substantially pure protein or polypeptide is isolated from transfected or transformed cells containing an expression vector encoding the BAP28 protein or a portion thereof. The concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few micrograms/ml. Monoclonal or polyclonal antibody to the protein can then be prepared as follows:

A. Monoclonal Antibody Production by Hybridoma Fusion

Monoclonal antibody to epitopes in the BAP28 protein or a portion thereof can be prepared from murine hybridomas according to the classical method of Kohler, G. and Milstein, C., Nature 256:495 (1975) or derivative methods thereof. Also see Harlow, E., and D. Lane. 1988. Antibodies A Laboratory Manual. Cold Spring Harbor Laboratory. pp. 53-242.

Briefly, a mouse is repetitively inoculated with a few micrograms of the BAP28 protein or a portion thereof over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as ELISA, as originally described by Engvall, (1980), and derivative methods thereof. Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. et al. Basic Methods in Molecular Biology Elsevier, New York. Section 21-2.

B. Polyclonal Antibody Production by Immunization

Polyclonal antiserum containing antibodies to heterogeneous epitopes in the BAP28 protein or a portion thereof can be prepared by immunizing suitable non-human animal with the BAP28 protein or a portion thereof, which can be unmodified or modified to enhance immunogenicity. A suitable non-human animal is preferably a non-human mammal is selected, usually a mouse, rat, rabbit, goat, or horse. Alternatively, a crude preparation which has been enriched for BAP28 concentration can be used to generate antibodies. Such proteins, fragments or preparations are introduced into the non-human mammal in the presence of an appropriate adjuvant (e.g. aluminum hydroxide, RIBI, etc.) which is known in the art. In addition the protein, fragment or preparation can be pretreated with an agent which will increase antigenicity, such agents are known in the art and include, for example, methylated bovine serum albumin (mBSA), bovine serum albumin (BSA), Hepatitis B surface antigen, and keyhole limpet hemocyanin (KLH). Serum from the immunized animal is collected, treated and tested according to known procedures. If the serum contains polyclonal antibodies to undesired epitopes, the polyclonal antibodies can be purified by immunoaffinity chromatography.

Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. Also, host animals vary in response to site of inoculations and dose, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. Techniques for producing and processing polyclonal antisera are known in the art, see for example, Mayer and Walker (1987). An effective immunization protocol for rabbits can be found in Vaitukaitis, J. et al. J. Clin. Endocrinol. Metab. 33:988-991 (1971).

Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall. See, for example, Ouchterlony, O. et al., (1973). Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 μM). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., Chap. 42 in: Manual of Clinical Immunology, 2d Ed. (Rose and Friedman, Eds.) Amer. Soc. For Microbiol., Washington, D.C. (1980).

Antibody preparations prepared according to either the monoclonal or the polyclonal protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample. The antibodies may also be used in therapeutic compositions for killing cells expressing the protein or reducing the levels of the protein in the body.

**Example 7**

**Tissular specificity of the *BAP28* expression.**

Synthesis of the cDNA

The mRNA used are human RNA from CLONTECH.

11.5 µl water treated with DEPC (diethyl pyrocarbonate) with 1 µl of human RNA (1 µg/µl) and 1 µl of oligo dT primer random (oligo dT hexamer) (20pmol/µl) were heated at 74°C for 2 min 30 s. Then the enzymatic mixture was added. The enzymatic mixture comprised 4µL 5X Reaction Buffer, 1µL dNTP 10mM each, 0.5µL Recombinant RNase Inibitor 40U/µL and 1µL MMLV Reverse Transcriptase 200U/µL. The sample was heated 1 h at 42°C, and 5 min at 94°C. Then 80 µl of water treated with DEPC were added. (kit Advantage RT-for-PCR.CLONTECH K1402-2) The synthezised cDNAs were stocked at – 20°C.

Amplification of the BAP28 amplicon

The cDNAs used in this experiment come from the cDNA preparation described above and from Marathon Ready cDNA from CLONTECH.

For each tissue, the following PCR reactions were done.

* First PCR reaction : The couple of primers used in this PCR was PCTAexALF12 (SEQ ID No 36)/ BAP283Ra6283 (SEQ ID No 32). There were located in exon A` and exon 43 of the BAP28 gene, respectively.

The PCR assay was performed using the following protocol :

| | |
|---|---|
| Final volume | 50 µl |
| Water | 19.8 µL |
| Buffer 3.3X | 15 µL |
| Mix dNTP (25mM each) | 4 µL |
| rttHXL PERKIN ELMER (2U/µL) | 1 µL |
| Primer PCTAexALF12 (20pmol/µL) | 1 µL |
| Primer BAP283Ra6283 (20pmol/µL) | 1 µL |
| cDNA | 6 µL |

After 3 min of denaturation, 2.2 µl of $Mg(OAc)_2$ 25 mM were added. The PCR was proceeded with 10 min at 94°C; 34 cycles of 30 sec at 94°C, and 3 min at 67°C; and 10 min at 72°C.

* Second PCR reaction (Nested PCR) : The couple of primers used in this PCR was PCTAexALF13n (SEQ ID No 37)/ BAP283Ra6324n (SEQ ID No 33). There were also located in exon A` and exon 43 of the BAP28 gene, respectively, and they were more dowstream than the first couple of primers.

The PCR assay was performed using the following protocol :

| | |
|---|---|
| Final volume | 50 µl |
| Water | 20.8 µL |
| Buffer 3.3X | 15 µL |
| Mix dNTP (25mM each) | 4 µL |
| rttHXL PERKIN ELMER (2U/µL) | 1 µL |
| Primer PCTAexALF13n (20pmol/µL) | 1 µL |
| Primer BAP283Ra6324n (20pmol/µL) | 1 µL |

Product of PCR N°1 5 µL

After 3 min of denaturation, 2.2 µl of $Mg(OAc)_2$ 25 mM were added. The PCR was proceeded with 10 min at 94°C; 34 cycles of 30 sec at 94°C, and 3 min at 67°C; and 10 min at 72°C.

The PCR products of the second PCR were analyzed on a 1% TAE1X gel.

The results are shown in Figure 13. The segment comprising the exons 43 to A has been observed in the following tissues : Marathon testis, Marathon hippocampus, Marathon leukemia (chronic myelogenous K-562), cDNA cerebellum, cDNA substantia nigra, cDNA thalamus, cDNA caudate nucleus, cDNA spinal cord, cDNA pitiutary gland and cDNA mammary gland.

In contract, this cDNA segment has not been observed in Marathon Brain, Marathon Cerebellum, Marathon Cerebral Cortex, Marathon Hypothalamus, Marathon Fetal Kidney, Marathon Thyroid, Marathon Bone Marrow, Marathon HL60, Marathon MOLT4, Marathon Fetal Liver, Marathon Stomach, Marathon Prostate, cDNA Testis, cDNA Corpus Callosum, cDNA Amygdala, cDNA Fetal Brain, cDNA Skeletal Muscle, cDNA Lung, cDNA Kidney, cDNA Placenta, cDNA Spleen, cDNA Fetal Liver, cDNA Thyroid Gland, cDNA MOLT4, cDNA Adrenal Gland, cDNA Trachea, cDNA Salivary Gland, cDNA HL60, cDNA Small Intestine, cDNA Pancreas, cDNA Stomach, cDNA Bone Marrow, cDNA Thymus, cDNA Uterus, and cDNA Prostate.

An additional analysis of the expression pattern in the tissue has been done by the search of ESTs in Genbank database which show homology with the BAP28 cDNA. The results are shown in Table 6.

**Table 6**

| Tissue | Accession number in Genbank |
|---|---|
| placenta | AK001857 ; AI277866 |
| colon | AW858897 ; AW858960 |
| colon tumor metastasis | AW962967 |
| HeLa cell | AA098827 |
| Adipose tissue white | AA320776 |
| LNCAP cells | AA357743 |
| Total fetus | AA424101 ; AA460031 ; AA992680 |
| germinal center B cell | AA814857 ; AA814859 |
| testis | AI023607 ; AL040338 ; AA437086 |
| Fetal liver spleen | AI033328 |
| Fetal liver | AI114709 |
| Fetal heart | AI150773 |
| lung | AI348668 ; AW450486 |
| kidney | AI582623 |
| colon tumor | AI738790 |
| pooled fetal lung testis B-cell | AI827817 |
| stomach | AW389900 |
| Multiple sclerosis | N77431 |
| fetal liver spleen | T85649 |
| anaplastic oligodendroglioma Organ: brain | AI356180 |
| breast | AI905672 |

**Example 8**

**Cloning of a BAP28 cDNA.**

We cloned the BAP28 cDNA consisting to the exons 1 to 45.

Synthesis of cDNAs

mRNAs were total human prostate RNA from CLONTECH (Lot N°8040072 – Ref Cat:64038).

11.5 µL water treated with DEPC with 1 µL Total Human Prostate RNA (1 µg/µL) and 1 µL primer oligodT BAP28polyTcourt (20pmol/µl) (ttttttttttttttgtata : SEQ ID No 57) were heated 2 min 30 sec at 74°C. Then the enzymatic mixture was added. The enzymatic mixture comprised 4 µL 5X Reaction Buffer, 1 µL mix dNTP10mM each, 0.5 µL Recombinant RNase Inibitor 40U/µL and 1 µL MMLV Reverse Transcriptase 200U/µL. The sample was heated 1 h at 42°C and 5 min at 94°C. Then, 80µl water treated with DEPC were added. The obtained cDNAs were stocked -20°C.

Amplification of the BAP28 segment to be cloned : (Double PCR Reaction)

A first PCR with a couple of primer BAP28LF12.1 (SEQ ID No 58) / BAP28LR6726.1 (SEQ ID No 59) was performed using the following protocol :

| | |
|---|---|
| Final volume | 50 µl |
| Water | 19.8 µL |
| Buffer 3.3X | 15 µL |
| Mix dNTP (25mM each) | 4 µL |
| rttHXL PERKIN ELMER (2U/µL) | 1 µL |
| Primer BAP28LF12.1 (20pmol/µL) | 1 µL |
| Primer BAP28LR6726.1 (20pmol/µL) | 1 µL |
| Preparation of cDNA | 6 µL |

After 3 min of denaturation, 2.2 µl of $Mg(OAc)_2$ 25 mM were added. The PCR was proceeded with 10 min at 94°C; 34 cycles of 30 sec at 94°C, and 8 min at 67°C; and 10 min at 72°C.

A second PCR reaction (Nested PCR) with a couple of primers BAP28LF26SalI (SEQ ID No 60) / BAP28LR6717SalI (SEQ ID No 61) was performed using the following protocol :

| | |
|---|---|
| Final volume | 50 µl |
| Water | 18.3 µL |
| Buffer 3.3X | 15 µL |
| Mix dNTP (25mM each) | 4 µL |
| VENT BIOLABS (2 U/µL) | 3.5 µL |
| Primer BAP28LF12.1 (20pmol/µL) | 1 µL |
| Primer BAP28LR6726.1 (20pmol/µL) | 1 µL |
| Product of PCR N°1 | 5 µL |

After 3 min of denaturation, 2.2 µl of $Mg(OAc)_2$ 25 mM were added. The PCR was proceeded with 10 min at 94°C; 34 cycles of 30 sec at 94°C, and 8 min at 67°C; and 10 min at 72°C.

As soon as the end of PCR, the phenol/chloform extraction was performed in order to avoid in degradation. Finally, the PCR product was precipitated with NaCl and ethanol.

The PCR product and the cloning vector pGEM11Zf(+) were both digested by the restriction endonuclease SalI. The digested vector was then dephosphorylated. The digested PCR product was ligated with the digested and dephosphorylated pGEM11Zf(+) vector. E.coli DH10B was transformed by the obtained vector and the bacteria containing the recombinant vector were selected. The positive clones contained an 6.8 kb insert which is the expected size for the entire BAP28 cDNA. The sequencing of the insert showed a cDNA consisting of the exons 1 to 45 of BAP28.

**Example 9**

**Natural antisense structure.**

The natural antisense structure observed in the BAP28 gene related to the PCTA 1 gene is conserved in the Drosophila. Indeed, the new CDS generated from the Genbank sequence AE00315 (gene CG10805) is located between the positions 97601 and 104127 of the sequence. Another CDS is described on the opposite strand as the gene CG10806. This CDS is located between the positions 107695 and 104389 of the sequence. Then, the distance between the two CDS is about 262 bp. Therefore, as the 3'UTR of the 2 genes are likely overlapping, the new gene gene CG10805 is a natural antisense of the gene CG10806 and the natural antisense organization of BAP28 is conserved in Drosphila.

While the preferred embodiment of the invention has been illustrated and described, it will be appreciated that various changes can be made therein by the one skilled in the art without departing from the spirit and scope of the invention.

**REFERENCES**

The following references cited herein are incorporated herein by reference in their entireties

Abbondanzo SJ et al., 1993, Methods in Enzymology, Academic Press, New York, pp. 803-823 / Ajioka R.S. et al., *Am. J. Hum. Genet.*, 60:1439-1447, 1997 / Altschul et al., 1990, J. Mol. Biol. 215(3):403-410 / Altschul et al., 1993, Nature Genetics 3:266-272 / Altschul et al., 1997, Nuc. Acids Res. 25:3389-3402 / Anton M. et al., 1995, J. Virol., 69 : 4600-4606 / Araki K et al. (1995) *Proc. Natl. Acad. Sci. USA.* 92(1):160-4. / Ausubel et al. (1989)Current Protocols in Molecular Biology, Green Publishing Associates and Wiley Interscience, N.Y. / Baubonis W. (1993) *Nucleic Acids Res.* 21(9):2025-9. / Beaucage et al., *Tetrahedron Lett* 1981, 22: 1859-1862 / Bowcock et al., 1998, International Patent Publication No WO98/12327) / Bradley A., 1987, Production and

analysis of chimaeric mice. *In*: E.J. Robertson (Ed.), Teratocarcinomas and embryonic stem cells: A practical approach. IRL Press, Oxford, pp.113. / Bram RJ et al., 1993, Mol. Cell Biol., **13** : 4760-4769 / Brown EL, Belagaje R, Ryan MJ, Khorana HG, *Methods Enzymol* 1979;68:109-151 / Brutlag et al. Comp. App. Biosci. 6:237-245, 1990 / Bush et al., 1997, J. Chromatogr., **777** : 311-328. / Capecchi, M.R. (1989a) *Science*, 244:1288-1292 / Capecchi, M.R. (1989b) *Trends Genet.*, 5:70-76 / Chai H. et al. (1993) *Biotechnol. Appl. Biochem.*18:259-273. / Chee et al. (1996) *Science*. 274:610-614. / Chen et al. (1987) *Mol. Cell. Biol.* 7:2745-2752. / Chen et al. *Proc. Natl. Acad. Sci. USA* 94/20 10756-10761,1997 / Chen and Kwok *Nucleic Acids Research* 25:347-353 1997 / Chen and Kwok *Nucleic Acids Research* 25:347-353 1997 / Cho RJ et al., 1998, Proc. Natl. Acad. Sci. USA, **95**(7) : 3752-3757. / Chou J.Y., 1989, Mol. Endocrinol., 3: 1511-1514. / Clark A.G. (1990) *Mol. Biol. Evol.* 7:111-122. / Coles R, Caswell R, Rubinsztein DC, *Hum Mol Genet* 1998;7:791-800 / Compton J. (1991) *Nature*. 350(6313):91-92. / Davis L.G., M.D. Dibner, and J.F. Battey, Basic Methods in Molecular Biology, ed., Elsevier Press, NY, 1986 / Dempster et al., (1977) *J. R. Stat. Soc.*, 39B:1-38. / Dent DS & Latchman DS (1993) The DNA mobility shift assay. In: *Transcription Factors: A Practical Approach* (Latchman DS, ed.) pp1-26. Oxford: IRL Press / Eckner R. et al. (1991) *EMBO J.* 10:3513-3522. / Edwards et Leatherbarrow, *Analytical Biochemistry*, 246, 1-6 (1997) / Engvall, E., Meth. Enzymol. 70:419 (1980) / Excoffier L. and Slatkin M. (1995) *Mol. Biol. Evol.*, 12(5): 921-927. / Feldman and Steg, 1996, Medecine/Sciences, synthese, 12:47-55 / Felici F., 1991, J. Mol. Biol., Vol. 222:301-310 / Fields and Song, 1989, Nature, **340** : 245-246 / Fisher, D., Chap. 42 in: Manual of Clinical Immunology, 2d Ed. (Rose and Friedman, Eds.) Amer. Soc. For Microbiol., Washington, D.C. (1980) / Flotte et al. (1992) *Am. J. Respir. Cell Mol. Biol.* 7:349-356. / Fodor et al. (1991) *Science* 251:767-777. / Fraley et al. (1979) *Proc. Natl. Acad. Sci. USA*. 76:3348-3352. / Fried M, Crothers DM, *Nucleic Acids Res* 1981;**9**:6505-6525 / Fromont-Racine M. et al., 1997, Nature Genetics, **16**(3) : 277-282. / Fuller S. A. et al. (1996) *Immunology in Current Protocols in Molecular Biology*, Ausubel et al.Eds, John Wiley & Sons, Inc., USA. / Furth P.A. et al. (1994) *Proc. Natl. Acad. Sci USA*. 91:9302-9306. / Garner MM, Revzin A, *Nucleic Acids Res* 1981;**9**:3047-3060 / Geysen H. Mario et al. 1984. Proc. Natl. Acad. Sci. U.S.A. 81:3998-4002 / Ghosh and Bacchawat, 1991, *Targeting of liposomes to hepatocytes*, IN: *Liver Diseases, Targeted diagnosis and therapy using specific receptors and ligands*. Wu et al. Eds., Marcel Dekeker, New York, pp. 87-104. / Gonnet et al., 1992, Science 256:1443-1445 / Gopal (1985) *Mol. Cell. Biol.*, 5:1188-1190. / Gossen M. et al. (1992) *Proc. Natl. Acad. Sci. USA*. 89:5547-5551. / Gossen M. et al. (1995) *Science*. 268:1766-1769. / Graham et al. (1973) *Virology* 52:456-457. / Green et al., *Ann. Rev. Biochem.* **55**:569-597 (1986) / Grompe, M. (1993) *Nature Genetics*. 5:111-117. / Grompe, M. et al. (1989)

*Proc. Natl. Acad. Sci. U.S.A.* 86:5855-5892. / Gu H. et al. (1993) *Cell* 73:1155-1164. / Gu H. et al. (1994) *Science* 265:103-106. / Guatelli J C et al. *Proc. Natl. Acad. Sci. USA.* 35:273-286. / Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS, *Nat Genet* 1996;**14**(4):441-447 / Haff L. A. and Smirnov I. P. (1997) *Genome Research*, 7:378-388. / Hames B.D. and Higgins S.J. (1985) *Nucleic Acid Hybridization: A Practical Approach.* Hames and Higgins Ed., IRL Press, Oxford. / Harju L, Weber T, Alexandrova L, Lukin M, Ranki M, Jalanko A, *Clin Chem* 1993;**39**(11Pt 1):2282-2287 / Harland et al. (1985) *J. Cell. Biol.* 101:1094-1095. / Harlow, E., and D. Lane. 1988. Antibodies A Laboratory Manual. Cold Spring Harbor Laboratory. pp. 53-242 / Harper JW et al., 1993, Cell, **75** : 805-816 / Hawley M.E. et al. (1994) *Am. J. Phys. Anthropol.* 18:104. / Henikoff and Henikoff, 1993, Proteins 17:49-61 / Higgins et al., 1996, Methods Enzymol. 266:383-402 / Hillier L. and Green P. *Methods Appl.*, 1991, **1**: 124-8. / Hoess et al. (1986) *Nucleic Acids Res.* 14:2287-2300. / Huang L. et al. (1996) *Cancer Res* 56(5):1137-1141. / Huygen et al. (1996) *Nature Medicine.* 2(8):893-898. / Izant JG, Weintraub H, *Cell* 1984 Apr;**36**(4):1007-15 / Julan et al. (1992) *J. Gen. Virol.* 73:3251-3255. / Kanegae Y. et al., *Nucl. Acids Res.* 23:3816-3821. / Karlin and Altschul, 1990, Proc. Natl. Acad. Sci. USA 87:2267-2268 / Khoury J. et al., *Fundamentals of Genetic Epidemiology, Oxford University Press, NY*, 1993 / Kim U-J. et al. (1996) *Genomics* 34:213-218. / Klein et al. (1987) *Nature*. 327:70-73. / Kohler, G. and Milstein, C., Nature 256:495 (1975) / Koller et al. (1992) *Annu. Rev. Immunol.* 10:705-730. / Kozal MJ, Shah N, Shen N, Yang R, Fucini R, Merigan TC, Richman DD, Morris D, Hubbell E, Chee M, Gingeras TR, Nat Med 1996;2(7):753-759 / Landegren U. et al. (1998) *Genome Research*, 8:769-776. / Lander and Schork, *Science*, 265, 2037-2048, 1994 / Lange K. (1997) *Mathematical and Statistical Methods for Genetic Analysis*. Springer, New York. / Lenhard T. et al. (1996) *Gene.* 169:187-190. / Linton M.F. et al. (1993) *J. Clin. Invest.* 92:3029-3037. / Liu Z. et al. (1994) *Proc. Natl. Acad. Sci. USA.* 91: 4528-4262. / Livak et al., *Nature Genetics*, 9:341-342, 1995 / Livak KJ, Hainer JW, *Hum Mutat* 1994;**3**(4):379-385 / Lockhart et al. *Nature Biotechnology* **14**: 1675-1680, 1996 / Lucas A.H., 1994, In : Development and Clinical Uses of Haempophilus b Conjugate: / Mackey K, Steinkamp A, Chomczynski P, 1998, *Mol Biotechmol*, 9(1):1-5 / Mansour S.L. et al. (1988) *Nature.* 336:348-352. / Marshall R. L. et al. (1994) *PCR Methods and Applications.* 4:80-84. / McCormick et al. (1994) *Genet. Anal. Tech. Appl.* 11:158-164. / McLaughlin B.A. et al. (1996) *Am. J. Hum. Genet.* 59:561-569. / Morton N.E., *Am.J. Hum.Genet.*, 7:277-318, 1955 / Muzyczka et al. (1992) *Curr. Topics in Micro. and Immunol.* 158.97-129. / Nada S. et al. (1993) *Cell* 73:1125-1135. / Nagy A. et al., 1993, Proc. Natl. Acad. Sci. USA, 90: 8424-8428. / Nangaku M. (1994) *Cell* 79, 1209-1220 / Narang SA, Hsiung HM, Brousseau R, *Methods Enzymol* 1979;68:90-98 / Neda et al. (1991) *J. Biol. Chem.* 266:14143-14146.

/ Newton et al. (1989) *Nucleic Acids Res.* 17:2503-2516. / Nickerson D.A. et al. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87:8923-8927. / Nicolau C. et al., 1987, Methods Enzymol., 149:157-76. / Nicolau et al. (1982) *Biochim. Biophys. Acta.* 721:185-190. / Nyren P, Pettersson B, Uhlen M, *Anal Biochem* 1993;**208**(1):171-175 / O'Reilly et al. (1992) *Baculovirus Expression Vectors: A Laboratory Manual.* W. H. Freeman and Co., New York. / Ohno et al. (1994) *Science.* 265:781-784. / Oldenburg K.R. et al., 1992, Proc. Natl. Acad. Sci., **89**:5393-5397. / Orita et al. (1989) *Proc. Natl. Acad. Sci. U.S.A.*86: 2776-2770. / Ott J., *Analysis of Human Genetic Linkage, John Hopkins University Press, Baltimore,* 1991 / Ouchterlony, O. et al., Chap. 19 in: Handbook of Experimental Immunology D. Wier (ed) Blackwell (1973) / Parmley and Smith, Gene, 1988, **73**:305-318 / Pastinen et al., *Genome Research* 1997; **7**:606-614 / Pearson and Lipman, 1988, Proc. Natl. Acad. Sci. USA 85(8):2444-2448 / Pease S. ans William R.S., 1990. Exp. Cell. Res., 190: 209-211. / Perlin et al. (1994) *Am. J. Hum. Genet.* 55:777-787. / Peterson et al., 1993, Proc. Natl. Acad. Sci. USA, **90** : 7593-7597. / Pietu et al. *Genome Research* **6**:492-503, 1996 / Potter et al. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81(22):7161-7165. / Ramunsen et al., 1997, Electrophoresis, **18** : 588-598. / Reid L.H. et al. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87:4299-4303. / Risch, N. and Merikangas, K. (*Science*, 273:1516-1517, 1996 / Robertson E., 1987, Embryo-derived stem cell lines. In: E.J. Robertson Ed. *Teratocarcinomas and embrionic stem cells: a practical approach.* IRL Press, Oxford, pp. 71. / Rossi et al., *Pharmacol. Ther.* **50**:245-254, (1991) / Roth J.A. et al. (1996) *Nature Medicine.* 2(9):985-991. / Roux et al. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86:9079-9083. / Ruano et al. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87:6296-6300. / Sambrook, J., Fritsch, E.F., and T. Maniatis. (1989) *Molecular Cloning: A Laboratory Manual.* 2ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. / Samson M, et al. (1996) *Nature*, 382(6593):722-725. / Samulski et al. (1989) *J. Virol.* 63:3822-3828. / Sanchez-Pescador R. (1988) *J. Clin. Microbiol.* 26(10):1934-1938. / Sarkar, G. and Sommer S.S. (1991) *Biotechniques.* / Sauer B. et al. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85:5166-5170. / Schaid D.J. et al., *Genet. Epidemiol.*,13:423-450, 1996 / Schedl A. et al., 1993a, Nature, 362: 258-261. / Schedl et al., 1993b, Nucleic Acids Res., 21: 4783-4787. / Schena et al. *Science* **270**:467-470, 1995 / Schena et al., 1996, Proc Natl Acad Sci U S A.,**93**(20):10614-10619. / Schneider et al.(1997) *Arlequin: A Software For Population Genetics Data Analysis.* University of Geneva. / Schwartz and Dayhoff, eds., 1978, Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure, Washington: National Biomedical Research Foundation / Sczakiel G. et al. (1995) *Trends Microbiol.* 3(6):213-217. / Shay J.W. et al., 1991, Biochem. Biophys. Acta, 1072: 1-7. / Sheffield, V.C. et al. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 49:699-706. / Shizuya et al. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89:8794-8797. / Shoemaker

DD, et al., *Nat Genet* 1996;**14**(4):450-456 / Smith (1957) *Ann. Hum. Genet.* 21:254-276. / Smith et al. (1983) *Mol. Cell. Biol.* 3:2156-2165. / Sosnowski RG, et al., *Proc Natl Acad Sci U S A* 1997;**94**:1119-1123 / Spielmann S. and Ewens W.J., *Am. J. Hum. Genet.*, 62:450-458, 1998 / Spielmann S. et al., *Am. J. Hum. Genet.*, 52:506-516, 1993 / Sternberg N.L. (1994) *Mamm. Genome.* 5:397-404. / Sternberg N.L. (1992) Trends Genet. 8:1-16. / Stryer, L., *Biochemistry*, 4th edition, 1995 / Syvanen AC, *Clin Chim Acta* 1994;**226**(2):225-236 / Szabo A. et al. *Curr Opin Struct Biol* **5**, 699-705 (1995) / Tacson et al. (1996) *Nature Medicine*. 2(8):888-892. / Tatusov, R. L. & Koonin, E. V. (1994), CABIOS 10, No 4 / Te Riele et al. (1990) Nature. 348:649-651. / Terwilliger J.D. and Ott J., *Handbook of Human Genetic Linkage*, *John Hopkins University Press*, *London*, 1994 / Thomas K.R. et al. (1986) *Cell.* 44:419-428. / Thomas K.R. et al. (1987) *Cell.* 51:503-512. / Thompson et al., 1994, Nucleic Acids Res. 22(2):4673-4680 / Tur-Kaspa et al. (1986) *Mol. Cell. Biol.* 6:716-718. / Tyagi et al. (1998) *Nature Biotechnology*. 16:49-53. / Urdea M.S. (1988) *Nucleic Acids Research.* 11:4937-4957. / Urdea M.S. et al.(1991) *Nucleic Acids Symp. Ser.* 24:197-200. / Vaitukaitis, J. et al. J. Clin. Endocrinol. Metab. 33:988-991 (1971) / Valadon P., et al., 1996, J. Mol. Biol., **261**:11-22. / Van der Lugt et al. (1991) *Gene*. 105:263-267. / Vanhee-Brossollet and Vaquero, (1998) Gene :211(1):1-9 / Vlasak R. et al. (1983) *Eur. J. Biochem.* 135:123-126. / Von Heijne, G., J. Mol. Biol. (1992) 225, 487-494 / Wabiko et al. (1986) *DNA*.5(4):305-314. / Walker et al. (1996) *Clin. Chem.* 42:9-13. / Wang et al., 1997, Chromatographia, **44** : 205-208. / Weir, B.S. (1996) *Genetic data Analysis II: Methods for Discrete population genetic Data*, *Sinauer Assoc., Inc., Sunderland, MA, U.S.A.* / Westerink M.A.J., 1995, Proc. Natl. Acad. Sci., **92**:4021-4025 / White, M.B. et al. (1992) *Genomics.* 12:301-306. / White, M.B. et al. (1997) *Genomics.* 12:301-306. / Wong et al. (1980) *Gene*. 10:87-94. / Wood S.A. et al., 1993, Proc. Natl. Acad. Sci. USA, 90: 4582-4585. / Wu and Wu (1987) *J. Biol. Chem.* 262:4429-4432. / Wu and Wu (1988) *Biochemistry*. 27:887-892. / Wu et al. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86:2757. / Yagi T. et al. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87:9918-9922. / Zhao et al., *Am. J. Hum. Genet.*, 63:225-240, 1998 / Zou Y. R. et al. (1994) *Curr. Biol.* 4:1099-1103.